Water End Use Clustering Using Hybrid Pattern Recognition Techniques — Artificial Bee Colony, Dynamic Time Warping and K-Medoids Clustering

A. Yang, H. Zhang, R. A. Stewart, and K. A. Nguyen

***Abstract*—The smart water meter collected data has made a great progress for the categorization of residential water end use events, the efficiency and accuracy still need to be improved. In this paper, an advanced algorithm is proposed for clustering the end-use category of a mechanical appliance. For this study, the database of end use events was collected using smart meters from over 200 households located in South-east Queensland (SEQ), Australia. Firstly, the raw data is pre-processed and physical characteristics (e.g., volume, duration, max flowrate, etc.) are extracted. Due to the type of the dataset is water end used flow data, which based on time series, a K-Medoids clustering algorithm based on the Dynamic Time Warping algorithm is used for clustering. In addition, a swarm intelligence which is named Artificial Bee Colony algorithm brings the whole system into equilibrium. Numerical experiments are based on toilet flushing events. Results indicate that the hybrid technique improves the clustering accuracy from 82.85% to 95.71%, and it can be implemented to other mechanical water end use events such as clothes washers and dish washers.**

***Index Terms*—Artificial bee colony algorithm, dynamic time warping algorithm, water end-use, K-Mediods clustering.**

I. Introduction

Pattern recognition has been a popular research field over the past few decades, such as character recognition, speech recognition and medical applications. With the advanced development of smart water meter technology, extensive water end use data has been collected and employed for various studies since 1998 such as USE Investigation of domestic water end use, Yarra Valley Water Residential End Use study and WA Water Corporation Domestic Water Use study. Most recently, Nguyen [1] developed an integrated intelligent pattern recognition model to automate the categorisation of residential water end-use events.

For most of existing water metering systems, the collected data, usually in quarterly basis, cannot provide real-time information or other management information about water service. To overcome this issue, a smart water management system based on smart metering technology was proposed which would provide real-time information that showed the how, when and where water is consumed for both user and the water utility [1]. Nguyen [2] applied Dynamic Time Warping (DTW) algorithm to categorise the water end use events. And one of the major parts of water end use classification task is water event clustering. In this part of research, if existing algorithms such as simple DTW algorithm, simple K-Medoids clustering and Artificial Neural Network (ANN) were applied to clustering water events, the accuracy and the efficiency are not enough to meet the needs of both users and water utilities. Nguyen [3] introduced an intelligent autonomous system for residential water end use classification which combines Hidden Markov Model (HMM), Artificial Neural Networks (ANN) and the Dynamic Time Warping (DTW) algorithm.

The purpose of this study is to develop an improved intelligent technique to cluster similar water end use events together. This hybrid technique consists of Artificial Bee Colony (ABC), Dynamic Time Warping algorithm (DTW) and K-Medoids Clustering. Its advantages include: (i) the ABC algorithm combines with K-Medoids Clustering, which has strong local search capability, could improve the comprehensive performance of clustering algorithm, (ii) the K-Medoids clustering based on DTW distance can group the mixed water events into different clusters according to the pattern characteristics, and (iii) the hybrid algorithm has the ability of global searching and optimization.

II. Intelligent Techniques for Time Series Clustering

A. Overview of Time Series Clustering Techniques

Rani and Sikka [4] stated that time-series clustering is one of the concepts of data mining and listed some techniques that are used in time series clustering and depend on distance measuring. For distance measuring, the Euclidean and DTW are two popular techniques. For times series clustering, K-Medoids is the most frequently occurring algorithm. However, these approaches have not yet been applied in water end use analysis. Han [5] found that clustering is a process of a set of objects moving into clusters and the characteristic is objects which in one cluster are similar but they are dissimilar to objects from other clusters. There are many algorithms for K-Medoids clustering but the partitioning around medoids (PAM) which was proposed by Kaufman and Rousseeuw [6] is the most popular approach. However, Han [5] pointed out that the PAM algorithm needs a long computational time for a large data set, so the

Manuscript received June 15, 2018; revised August 7, 2018..

A. Yang, H. Zhang, and R. A. Stewart are with the Griffith School of Engineering, Griffith University, QLD 4222, Australia (e-mail: ao.yang@griffithuni.edu.au, hong.zhang@griffith.edu.au, r.stewart@griffith.edu.au).

K. A. Nguyen is with the Cities Research Institute, Griffith University, QLD 4222, Australia (e-mail: k.nguyen@griffith.edu.au).

doi: 10.18178/ijmlc.2018.8.5.733

efficiency is low. However, Park and Jun [7] reported a new algorithm of K-medoids clustering which is simple and fast and has been tested in many different type of data sets. K-Medoids clustering is one of the centroid-based clustering or rather, an improved algorithm of K-means. Its advantages include: (i) K- Medoids algorithm has the ability to process a large dataset. (ii) K-Medoids algorithm can reduce the effect of the outliers on the results. However, K-Medoids clustering also has advantages such as the influence of initial medoids being strong and the ability of global searching being poor.

B. Data Clustering in Water End Use Analysis

DTW algorithm is a method for measuring the similarity between two time series with different lengths. In this task, each water end- use event can be defined as a times sequence and the objective is to find the distance between two events. Nguyen [1], [2] demonstrated an application of DTW algorithm for prototype selection and similar event clustering in the water end use classification task. However, the disadvantage of this technique was the heavy dependence on the use of threshold value, which defined the similarity between two samples. A large threshold value would allow two significantly different samples to be considered similar and clustered into the same group, while a small threshold value may assign two relatively similar samples into two different groups. As a consequence, different threshold value settings will result in different number of events in each cluster. The different number of events which are considered to be similar based on two different threshold value settings is shown in Fig. 1.

III. Hybrid Pattern Recognition Techniques

To overcome the disadvantage of DTW, a combination of K- Medoids, DTW algorithm and Artificial Bee Colony algorithm have been proposed where each approach will be in charge of a significant role in the whole algorithm. Presented below is a brief summary of each technique.

A. Dynamic Time Warping Algorithm

The main application of DTW algorithm is to determine the distance between two events which have a similar pattern. The processes of this algorithm include following three steps:

Step 1: Suppose that there are two time sequences of the water flow data $\mathbf{Q} = (q_1, q_2, \dots q_i, \dots q_m)$ and $\mathbf{C} = (c_1, c_2, \dots c_j, \dots c_n)$ of length m and n respectively. Define $d(q_i, c_j) = |q_i - c_j|$, and $\mathrm{D}(q_i, c_j)$ as the total DTW distance between $q(1{:}\mathrm{i})$ and $c(1{:}\mathrm{j})$ with warping path from (1,1) to (i, j), which follows $(1 \le i \le m)$ and $(1 \le j \le n)$.

Step 2: Accumulated distance using DTW algorithm as presented below:

$$D(q_i, c_j) = d(q_i, c_j) + \min D \quad (1)$$

The $minD$ represents the minimum distance between two nodes in warping path and written as:

$$minD = \{\mathrm{D}(q_{i-1}, c_j), \mathrm{D}(q_{i-1}, c_{j-1}), \mathrm{D}(q_i, c_{j-1})\} \text{ and } 1 \le i \le m \text{ and } 1 \le j \le n$$

The mapping path is from (q_1, c_1) to (q_m, c_n) and $D(q_1, c_1) = d(q_1, c_1)$ are the initial conditions of this algorithm. In addition, the Fig. 2 as shown below presents 3 directions in warping path and explains the Equation (1) clearly.

Step 3: The final Dynamic Time Warping distance between $\mathbf{Q}$ and $\mathbf{C}$ is $\mathrm{D}(q_m, c_n)$.

In this study, the main objective of this algorithm is to determine the distance between two different water end-use events and this distance as a benchmark in K-Medoids clustering.

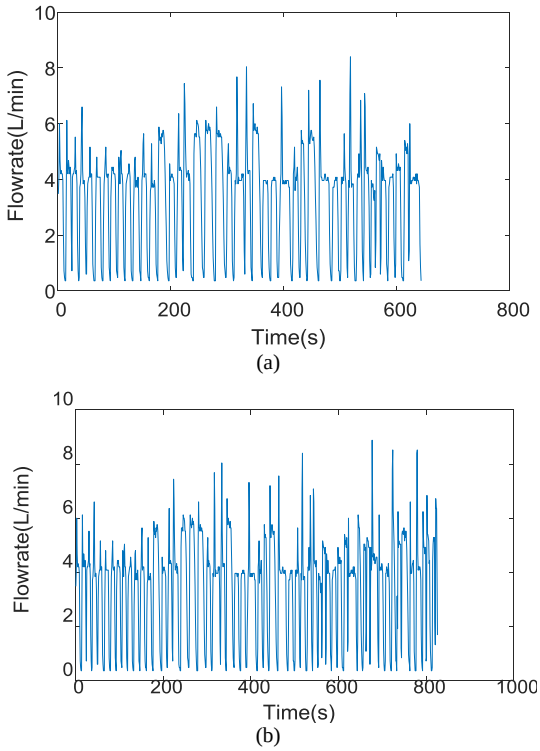


Fig. 1. Result of DTW algorithm ((a) Threshold value=700, Identified events=43. (b) Threshold value=350, Identified events=56).

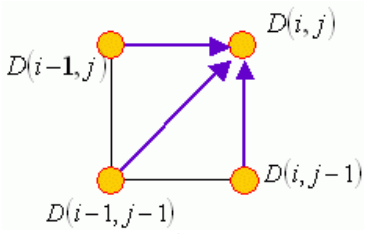


Fig. 2. Three directions of grid node in warping path.

B. K-Medoids Clustering

As an unsupervised learning, clustering plays an important role in pattern recognition. In this study, the dataset is the flowrate time series, and distance measurement is based on DTW distance. Here, the K-Medoids clustering is divided by five steps:

Step 1: Randomly select K samples as the initial cluster centroid from n objects which represents by $C_1, C_2, \dots C_j, \dots C_k$.

Step 2: Determine the DTW distance between each object and medoids following the procedure in Section III.A. Here, the equation of DTW distance is written by:

$$d(ij) = dtw\{O(i), C(j)\} \quad 1 \le i \le n \; and \; 1 \le j \le k \quad (2)$$

where $d(ij)$ represents the DTW distance between object i and medoid j.

Step 3: Assign objects to each cluster according to the

minimum distance principle. Such as $O(i) \in \text{Group}\{min\{d(ij)\}\}$.

Step 4: Select each object as new cluster medoid $C_i (1 \le i \le n)$ orderly from each cluster and find the optimal solution. In this algorithm, the sum of DTW distance between objects and medoid is the criteria of the optimal result. The new cluster centroid C_i which have minimum result of cost function will replace the initial cluster centroid C_j.

Step 5: Repeat step 2 to step 4 until the cluster centroid does not change.

C. Artificial Bee Colony (ABC) Algorithm

ABC algorithm is an optimisation technique based on the foraging behaviour of honey bee in swarm intelligence. Karaboga [8] proposed the ABC algorithm, and in the beginning, applied it in solving function. Then, Karaboga [9], [10] pointed out that compared with other swarm intelligence such as particle swarm optimization (PSO) and evolutionary algorithm (EA), the ABC algorithm has higher efficiency. The specific description of this algorithm is: employed bees onto food source and sharing the information with onlookers, if there is a food source abandoned by employed bees or onlookers, the employed bee of this food source will turn into a scout. The mission of scouts is to find the new food source. According to this theory, the ABC algorithm can be divided into 4 steps:

Step 1: Initial Population: A random bee colony is initialised, $B = \{B_1, B_2, \dots, B_i, \dots B_{SN}\}$ which has SN solutions (food sources). Each solution $B_i = \{b_{i1}, b_{i2}, \dots, b_{iD}\}$ is a vector with D-dimension and D represents the number of parameters. The food source is randomly generated by the equation:

$$B_i^j = B_{min}^j + rand(0,1)\left(B_{max}^j - B_{min}^j\right) \; j \in \{1, 2, \dots, D\} \quad (3)$$

where, B_i^j is the jth dimension of the solution vector;

B_{max}^j is the maximum value of the jth dimension;

B_{min}^j is the minimum value of the jth dimension;

$rand(0,1)$ is the uniform random number in interval (0,1).

In addition, the maximum cycle number is defined as N_{max}.

Step 2: Employed Phase: In this step, employed bees randomly select a new position of food source around the old one and the equation of search new food source is:

$$V_i^j = B_i^j + R_i^j\left(B_i^j - B_k^j\right) \; j \in \{1, 2, \dots, D\} \; and \; k \in \{1, 2, \dots, SN\} \quad (4)$$

where, R_i^j is a random number between [-1, 1].

After greedy selection of new position of food source, employed bees will determine the fitness function, if the fitness function of the new source is higher than that of the previous one, the new position V_i^j will replace the B_i^j. Otherwise, bees will keep the old position B_i^j.

Step 3: Onlooker Phase: After all the employed bees complete the neighbour search, an onlooker bee will choose a food source depending on the probability value P_i, written by the following equation:

$$P_i = \frac{F_i}{\sum_{n=1}^{SN} F_n} \quad (5)$$

where SN is the number of food sources and F_i is the ftness value of solution *i* which represents the nectar amount of the food source. In addition, onlookers will also use greedy selection and follow Step 2 to choose food source in neighbour search.

Step 4: Scout Phase: In this step, if a position of food source cannot be updated within the *limit* times which parameters in order to abandon food source, this food source is assumed to be abandoned and the employed bee of this food source will change into a scout. In addition, the scout bee will search a new food source according to Equation (3) to replace the previous one. Repeating the above steps until cycle numbers meet the N_{max}.

D. Hybrid Algorithm Implementation to Water End Use Analysis

K-Medoids clustering is an algorithm which has advantages such as being simple, fast for convergence and robust for partly searching. However, it can be improved because the dependence on the initial medoid is strong and the ability of global research is poor. For this reason, a swarm intelligence needs to be used in this study to reduce the impact of the initial medoid on the combined algorithm. This section will explain how these three techniques have been integrated to improve the clustering efficiency. The basic steps of techniques are described below and the hybrid techniques operation flow chart is summarized in Fig. 3.

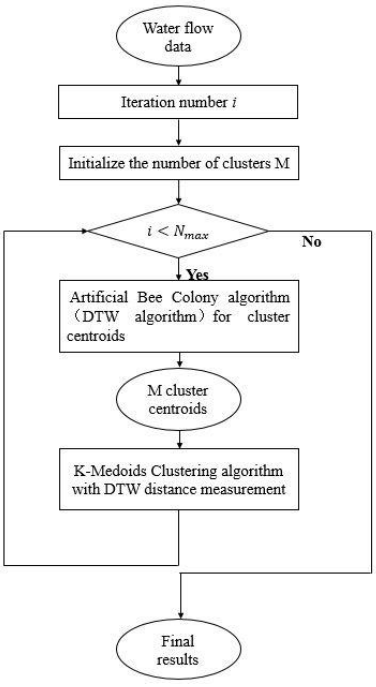


Fig. 3. Improved technique flow chart.

Step 1: The first required step of the hybrid algorithm is to initialize parameters and select the initial medoids. The number of bees is N_B, and the first half of the colony is employed bees and the second half consists of onlookers. The number of cluster is equal to M. The maximum iteration number and control parameter are N_{max} and *L* respectively.

Step 2: Each bee is employed to search food source using the Equation (4), then the probability P_i is calculated when all employed bees finish their searching. Every onlooker will choose their employed bees according to the P_i and employed bees will find a new source again.

Step 3: If the fitness function a food source is not the best result for the whole colony after L iteration, the employed bee will change into a scout and find a new source using the Equation (3). In this study, the fitness function is written by:

$$F_i = \sum_{j=1}^{M} \sum_{x \in C_j} DTW(x, C_j) \quad (6)$$

where F_i represents the sum of DTW distance between objects, x and medoid of the j^{th} cluster, C_j in i^{th} iteration.

Step 4: Using the result of Step1 as the initial medoid and operate a K-Medoids clustering for the dataset. A new group of medoids will be calculated from this procedure and use these medoids to update the bee colony. If the number of iterations is less than N_{max} , repeating the above steps. Otherwise, the hybrid algorithm will stop and get the optimal medoids.

Step 5: As the initial points, these optimal medoids are used in the K-Medoids clustering and each cluster can be determined.

IV. Experimental Results and Discussion

The water end use datasets utilised for hybrid algorithm were from the South-east Queensland Residential End Use Study (SEQREUS). The data is recorded by high-resolution smart water meter, which collect 0.014L/pulse water consumption data every 5 seconds, from over 200 homes and analysed by Trace Wizard™ [11] The mixed water events consist of nine different water categories, including shower, tap, dishwasher, clothes water, bathtub, irrigation, leak, full-flush toilet and half-flush toilet. In this experiment, a dataset of 136 water end use events, including 70 toilet events mixed with 66 events from other categories, was used to test the hybrid algorithm. The overall objective is to group all toilet events together, and the method efficiency is evaluated based on the number of toilet events grouped together over the number of total toilets event present in the dataset. For the Artificial Bee Colony algorithm in proposed algorithm, the parameters are set as follows, which are the colony size $N_B = 100$, the maximum cycles number $N_{max} = 500$ and the L is equal to 200.

E. Experimental Verification

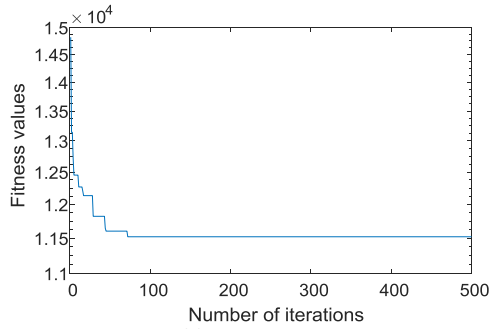


Fig. 4. Changes of fitness values in testing dataset.

The convergence graph of hybrid algorithm, which is based on the dataset, is shown in Fig. 4. K-Medoids clustering makes the algorithm quickly reach the local extremum value and the Artificial Bee Colony allows the algorithm to escape the local maxima and reach the global optimal value. The results will not change in a large number of experiments. This result also shows that this hybrid algorithm has a good effect in convergence.

The classified results of Dynamic Time Warping algorithm and hybrid techniques are shown in Fig. 5.

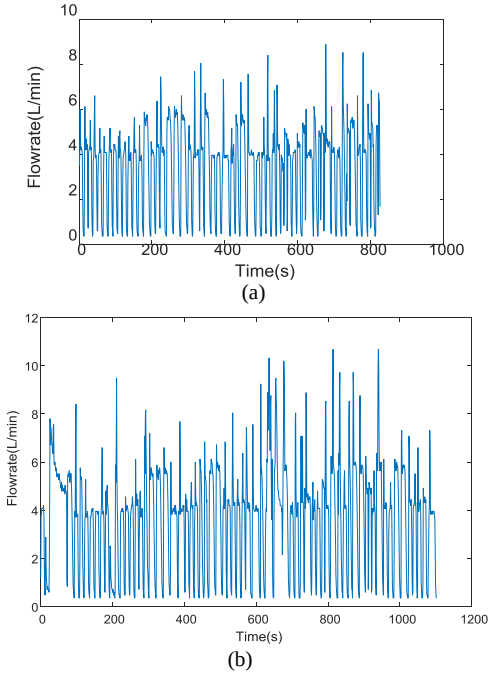


Fig. 5. (a): Toilet events categorized using Dynamic Time Warping algorithm; (b): Toilet events categorized using proposed hybrid algorithm.

Fig. 5(a) shows that 58 toilet events were grouped together using the DTW algorithm alone as presented in (Nguyen et al., 2011), while the proposed algorithm has successfully clustered 67 toilet events. The overall testing results was presented in Table I below.

TABLE I: Comparison of the Performance between 2 Different Algorithms

Techniques	Identified events	Total toilet events	Testing Accuracy
DTW algorithm	58	70	82.85%
Hybrid techniques	67	70	95.71%

F. Discussion

Compared to the Dynamic Time Warping algorithm, the hybrid algorithm is capable of maintaining the accuracy and also improving the efficiency. Due to combination of ABC algorithm, clustering results are relatively stable. When applying this technique, the distance from medoids to all other samples are required to be determined many times until the whole system is in equilibrium. The condition of equilibrium is the sum dynamic time warping distance of each cluster is the minimum, which represents the main theory of clustering.

V. Conclusion and Future Work

The study presented an improved pattern recognition technique in water end-use events analysis. This hybrid

algorithm can also be used as a part of an improved technique for improving the smart water metering analysis system. In addition, an out-of-order dataset can be classified and better analysed through this hybrid technique. In the testing dataset, the results of experiments show that this algorithm has high efficiency, accuracy, and practicability. However, the limitation of this method is that it still requires an initial selection of the number of clusters prior to running the algorithm, and if this number is set inappropriately, the final clustering result will be reduced. In conclusion, the proposed algorithm is recommended for all the problems which are based on time series analysis.

In the analysis of water end use, toilet event is a part of mechanical processes, clothes washer and dish washers are the same type end use as toilet flushing events and they have also a high degree of pattern and periodicity. The next task is to apply this hybrid algorithm in these two water end uses and other energy pattern analyses such electricity or natural gas.

REFERENCES

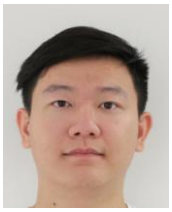
[1] K. A. Nguyen, R. A. Stewart, and H. Zhang, "An intelligent pattern recognition model to automate the categorisation of residential water end-use events," *Environmental Modelling & Software*, vol. 47, pp. 108-127, 2013.

[2] K. A. Nguyen, H. Zhang, and R. A. Stewart, "Application of dynamic time warping algorithm in prototype selection for the disaggregation of domestic water flow *idea based on honey bee swarm for numerical optimization*," Technical report-tr06, Erciyes University, Engineering faculty, Computer Engineering Department, vol. 200, 2005.

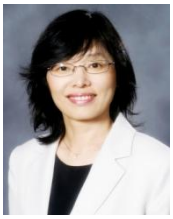
[3] D. Karaboga and B. Basturk, "On the performance of artificial bee colony (ABC) algorithm,"*Applied Soft Computing*, vol. 8, no. 1, pp. 687-697, 2008.

[4] D. Karaboga and B. Basturk, "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm," *Journal of Global Optimization*, vol. 39, no. 3, pp. 459-471, 2007.

[5] K. A. Nguyen, H. Zhang, and R. A. Stewart, "Development of an intelligent model to categorise residential water end use events," *Journal of Hydro-Environment Research*, vol. 7, no. 3, pp. 182-201, 2013.



Ao Yang was born in Chifeng, Inner Mongolia Autonomous Region, China in 1994. He graduated from the School of Engineering and Built Environment, Griffith University, Gold Coast, Australia in 2016. Then he is an MPhil candidate in Griffith University. His recent research interest is in fields of water end-use study.



Hong Zhang is a professor at the School of Engineering and Built Environment, Griffith University. Professor Zhang has active research interests in the fields of water resource engineering (dynamics of groundwater flow, river sediment transport, water quality and dynamics in lakes and ponds, water end-use) and the coastal/ocean dynamics (circulations, wave dynamics, mixing processes, wave-structure-soil interactions). She employs a variety of techniques such as analytical, experimental, numerical and artificial neural network methods to obtain the theoretical understanding of her research problems and apply them to engineering practices.

Rodney Stewart is a professor at the School of Engineering and Built Environment, Griffith University. Professor Rodney Stewart is an expert in engineering, construction and environmental engineering and management research. His current particular area of research focus is on digital utility transformation. Professor Stewart is leading industry collaborative research projects that seek to integrate 'big data' metering and monitoring technologies and associated expert systems into infrastructure, particularly in the water and energy utility sector, in order to better manage these critical resources and better integrate contemporary solutions such as renewable energy and decentralized water supply.

Khoi Nguyen is an expert in engineering and computer science. His current research areas are water engineering and digital utility transformation. He employs a variety of pattern recognition and machine learning techniques to help disaggregate total water and energy consumption into a repository of end-use category.