Classification of excessive domestic water consumption using Fuzzy Clustering Method

A. Zairi Zaidi and Khairul A. Rasmani

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Seremban Campus, 70300 Seremban, Negeri Sembilan, Malaysia.

E-mail: ahmadzairi@gmail.com, khairulanwar@ns.uitm.edu.my

Abstract. Demand for clean and treated water is increasing all over the world. Therefore it is crucial to conserve water for better use and to avoid unnecessary, excessive consumption or wastage of this natural resource. Classification of excessive domestic water consumption is a difficult task due to the complexity in determining the amount of water usage per activity, especially as the data is known to vary between individuals. In this study, classification of excessive domestic water consumption is carried out using a well-known Fuzzy C-Means (FCM) clustering algorithm. Consumer data containing information on daily, weekly and monthly domestic water usage was employed for the purpose of classification. Using the same dataset, the result produced by the FCM clustering algorithm is compared with the result obtained from a statistical control chart. The finding of this study demonstrates the potential use of the FCM clustering algorithm for the classification of domestic consumer water consumption data.

1. Introduction

Water is an important natural resource to all living things. Every organism uses water for living needs and for other activities to fulfil their needs. This indicates that the greater the world population, the greater the demand for water. Nowadays, it can be seen that the increase in demand for water is not in accordance with the supply of water. There is insufficient water supply to meet the increasing demand in many places [1]. This shows that conservation of water is much needed in response to the water shortage. Therefore, efficient management of clean and treated water is necessary to avoid excessive consumption, which might lead to wastage of water for unnecessary purposes.

Classification of excessive domestic water consumption is crucial; the process involves the complexity that arises from uncertain information in the consumer data resulting from uncontrolled factors, such as the behaviour of the consumer in terms of socio-economic and socio-cultural aspects. Hence, it is expected that different individuals will have different behaviour regarding water usage. Besides that, the environmental factor is believed to make a significant contribution to the variation of per capita domestic water consumption for different localities. Therefore, it is difficult to determine excessive domestic water usage for each premise solely on the basis of per capita domestic consumption. Without a proper method to classify the data as ‘excessive’ or ‘non-excessive’ water usage, it is...
difficult to create a mathematical model using supervised learning methods. With the absence of class labels, there is great potential to employ an unsupervised learning technique for the study of the classification of excessive domestic water consumption. Unsupervised learning techniques are methods that can deal with unlabelled data, whereby the algorithm will carry out inferences and finally classify the data into homogenous classes.

The fuzzy clustering technique has been applied in many studies, such as the classification of coal seams [2], software quality prediction [3], classifying sets of oral cancer cell data [4] and market segmentation [5]. In this study, it is proposed that the classification of excessive domestic water consumption is conducted using the Fuzzy C-Means (FCM) clustering algorithm. Consumer data is used in this study instead of the duration and flow of water because the data collection process is easier and does not require a specific physical tool. FCM is an unsupervised learning algorithm that is used to group a data set into a number of clusters, whereby every data point has its membership degree of belongingness to every cluster [6].

2. Background

Appropriate clean water management, such as water distribution network management, the estimation and production of future water consumption and sophisticated water management systems are necessary to conserve water for current and future use. The complexity of water management and demand from consumers lead to the need for an effective water management system. Savic & Walters [7] suggested ‘hydroinformatics’ as an approach to water network management and maintenance. The researchers argued that the discipline of ‘hydroinformatics’ helps in the development and application of innovative technology for managing the water distribution network and they indicate three suitable methods that can be applied in the water industry: a geographical information system, artificial neural networks and a genetic algorithm. It has been claimed that the cost saving of using these methods ranges from 5%-50%. Makropoulos & Butler [8] discussed the nature of uncertainty in urban water management, which involves the determination of optimum locations for facilities in a particular area with respect to economic and environmental factors. It has been suggested that the type-1 and type-2 fuzzy inference system (FIS) can be useful tools to solve water management problems with uncertainties and can also be adopted for spatial decision support systems (SDSS) in urban water management.

Water consumption prediction and estimation plays a crucial part in treated water management. It helps in forecasting water consumption for better management of clean water in the future. Naizhuo & Yunsheng [9] conducted a study to estimate domestic water consumption in China and the United States at province/state level using night-time satellite imagery. Since a large population leads to high cost and difficulty in collecting data by traditional statistical methods, the study aims to investigate the potential of using night-time satellite imagery where the sum light derived from the imagery data is compared with the actual water consumption data. The result of the study shows that the proposed method has potential for the estimation of domestic water consumption at national or sub-national level in the future.

Altunkaynak, Özger, & Çakmakci, [10] proposed the use of a fuzzy model to predict future water consumption for Istanbul city. The Adaptive Neuro Fuzzy Inference System (ANFIS) was employed to estimate the membership function parameters and the Takagi-Segeno Fuzzy Inference System was used to generate the output. With overall relative error of less than 10%, the model can be considered as practically acceptable to predict water consumption in Istanbul city. Şen & Altunkaynak [11] investigated the use of a fuzzy approach in a different scenario, whereby a fuzzy model was established for predicting individual drinking water consumption. The researchers argued that many statistical approaches lead to loss of information or miscalculation since they involve crisp interval values, whereas statistical assumption made for regression analysis may not frequently be achieved in
practical terms. Thus, a fuzzy model based on the Mamdani approach was proposed in the study, in which vague data such as human weight, weather, temperature and human physical activity were utilised to develop the model. The proposed model was considered acceptable for practical individual drinking water consumption rate predictions since the average relative error is less than 5%.

Corona-Nakamura et al. [12] proposed a model to identify automatically the output of domestic water consumption based on duration and flow of water. It has been argued by the researchers that a suitable model could be easily interpreted by humans and can include information provided by users or experts. In the study, the researchers use the Adaptive-Neuro Fuzzy Inference System (ANFIS) to classify types of water consumption activity in residential premises. The finding of the research shows that the model can be a good solution to the problem: it has a good recognition percentage and small errors. Moreover, Corona-Nakamura, Ruelas, Ojeda-Magana, & Finch [13] conducted a research study to develop an unsupervised learning model that can automatically identify the point of water consumption in a house. The unsupervised learning model is beneficial since it can adapt to particular cases automatically. The researchers employed the Fuzzy C-Means (FCM) technique for estimating the class centre and an Improved Gustafson-Kessel algorithm for clustering of the data set. It was claimed that the developed model has a high recognition percentage and can be the basis of a supervisory system for water conservation.

It can be observed from the literature that, although much research has been conducted related to water consumption prediction, little research has focussed on the study of domestic water consumption at residential premises by utilising consumer data. Even though the consumer data includes uncertain information, especially on the quantity of water used for each activity related to individual water consumption, the collection of such data is relatively easier as it will not involve any specific measuring tools. Additionally, it can be seen that there is already research that employs a fuzzy clustering method for the classification of water consumption [12, 13]. But, this research employed data that was more complicated to collect by consumers or water supply companies. Therefore, consumer data can be explored to find its potential use for research on the classification of water consumption using a fuzzy clustering method. This motivates the present study that focuses on employing the Fuzzy C-Means (FCM) clustering algorithm for the classification of domestic consumer water consumption data.

3. Fuzzy Clustering Method

Data clustering is the process of grouping data into the same group: the data are similar to each other in the same group and dissimilar from the data in other groups [14]. Fuzzy clustering is one of the many existing data clustering techniques. In hard clustering, data are grouped to exactly one cluster. On the other hand, in fuzzy clustering (or soft clustering), data can be grouped into more than several clusters where the data will have membership degree to each cluster in the interval [0, 1]; the larger the membership degree, the greater the association of the data to a particular cluster [15]. The Fuzzy C-Means (FCM) clustering algorithm [16] is the most commonly used algorithm in fuzzy clustering analysis [15]. This algorithm will group a data set into a known number of clusters. It aims to minimise iteratively the objective function $J(U, V)$, which is given by:

$$J(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij}^m \left\| x_i - v_j \right\|^2$$

where $n$ is the number of data points, $c$ is the number of clusters, $x_i$ is $i$th data point in the data set, $v_j$ is the $j$th cluster centre, $\mu_{ij}$ is the value of membership degree of $i$th data point in the $j$th cluster, $m\in[1, \infty)$ is the fuzziness index which determines the level of clustering result fuzzines [17], $U$ is the partition matrix, $V$ is the cluster centre matrix and $\left\| x_i - v_j \right\|$ is the distance between a data point $x_i$ and the cluster
centre $v_j$ and it is based on Euclidean distance. This algorithm will iterate until there is no further minimisation of the objective function by updating membership degree $u_{ij}$ and cluster centre $v_{ij}$.

There are several steps involved in the process of grouping data set using this FCM algorithm. Each iteration is labelled as $r$. The steps are explained as follows:

Step 1: Decide the number of clusters, $c$, where $2 \leq c < n$. In order to employ the FCM, the number of clusters must be known.

Step 2: Initialise partition matrix $U(r=0)$ randomly.

Step 3: Calculate the cluster centres matrix $V$ using the equation:

$$v_j = \frac{\sum_{i=1}^{n} u_{ij}^m x_i}{\sum_{i=1}^{n} u_{ij}^m}$$

(2)

Step 4: Calculate the distance matrix $D$ using the equation:

$$d_{ij} = \left\| x_i - v_j \right\|^2$$

(3)

Step 5: Update the partition matrix, $U(r+1)$

$$\mu_{ir} = \frac{1}{\sum_{r=1}^{c} \left( \frac{\left\| x_i - v_j \right\|}{m-1} \right)^2}$$

(4)

Step 6: If $\left\| U^{r+1} - U^r \right\| < \varepsilon$ where $\varepsilon$ is a termination criterion in $[0, 1]$ and $r$ are iteration steps, stop the iteration or else repeat Step 3 to update the cluster centres and the partition matrix $U$.

4. Experimental Setup

The purpose of this research is to classify consumer water consumption data using the Fuzzy C-Means (FCM) clustering algorithm. A data set from the study by Mohd. Hanif et al. [18] has been used in this study. It consists of 439 observations, which represent 439 residential premises. Eleven variables are considered in this study. They are the number of households, the number of toilets, the number of times teeth are brushed, the number of times hands and face are washed, the number of times a bath/shower is taken, the number of times the toilet is flushed, the number of times clothes are washed by hand, the number of times clothes are washed in the washing machine, the number of times cooking takes place, the number of times the car is washed and the average amount of monthly water usage (in litres).

In this study, the FCM clustering algorithm available in MATLAB® software was used. The algorithm initially guesses the cluster centres for the data set [6] and iteratively updates the cluster centres and membership degree for each data point until the objective function (Equation 1) is minimised. When the iteration is stopped, the cluster centres will be located in the right position, whereby each data point has membership degree to each cluster. The higher the membership degree, the greater the association of a data point to a cluster.
The algorithm will also create a matrix of cluster centres showing the coordinates of cluster centres for each data point, a partition matrix containing the membership degree for each data point in each cluster and the values of objective function for each iteration. In this study, the number of clusters, \( c \) was set to two which are “likely” and “unlikely” in order to indicate the excessiveness of water consumption. The fuzzier, \( m \) was set by default to two since there is no experimentation and domain knowledge of the value [15].

5. Findings and Discussion

Using the MATLAB® software, the FCM algorithm took 91 iterations to optimise the objective function. The result shows that, after the 91st iteration, no further process is made by the algorithm. This is because all data points have been grouped to their clusters with their own cluster centres and membership degree. At this point, the data points are associated with each cluster by membership degree. Figure 1 shows the example of the graph of the average amount of monthly water usage against a predictive variable. It can be seen that there are two partitions in the data set with their cluster centres marked with two symbols, ‘O’ and ‘×’. Such analysis is very useful to show that the FCM algorithm can be used to separate the data into different classes, which can be labelled as ‘likely’ and ‘unlikely’ cases of excessive water usage. This analysis is also very useful in demonstrating how FCM can be used to analyse how each predictive variable contributed to the final classification outcomes.

In order to validate the clustering outcome, the result produced by the FCM algorithm is compared with the result obtained from the study by Mohd Hanif et al. [18], where the Individual Hotelling T\(^2\) control chart was used. Table I shows the clustering result using FCM and the result from the statistical control chart, which shows that the similarity percentage between the classification outcomes is 86.33%.

![Figure 1. Example of clustering result for average monthly water usage against a predictive variable.](image)

| Method                        | Likely cases | Unlikely cases | Similarity Percentage |
|-------------------------------|--------------|----------------|-----------------------|
| Individual Hotelling T\(^2\) control chart | 41           | 398            |                       |
| Fuzzy C-Means clustering algorithm     | 47           | 392            | 86.33%                |
6. Conclusion
In this paper, it has been shown that the Fuzzy C-Means (FCM) clustering algorithm has been successfully employed to classify excessive domestic water consumption using consumer data. The result from this study was compared with other unsupervised learning method, which demonstrates the potential use of FCM for the classification of domestic consumer water consumption data. Note that, in this study, there is no validation data available. Hence, further research should look for the use of observed data collected from house visits to obtain class labels so that comparison and validation can be made with other supervised learning classification methods.

Acknowledgments
This research work is funded by the Ministry of Higher Education, Malaysia under the Fundamental Research Grant Scheme (FRGS) with reference number 600-RMI/FRGS TD 5/3 (1/2015). The authors also would like to thank the Institute of Research Management and Innovation (IRMI), Universiti Teknologi MARA, Malaysia.

References
[1] Arbués F, García-Valiñas M Á, and Martinez-Espiñeira R 2003 The Journal of Socio-Economics 32 81-102
[2] Sahu H B, Mahapatra S S, and Panigrahi D C 2012 Fuel Processing Technology 104 115-120
[3] Yuan X, Khoshgoftaar T M, Allen E B, and Ganesan K 2000 Proc. of the 3rd IEEE Symp. on Application-Specific Systems and Software Engineering Technology 85-90
[4] Wang X Y, Garibaldi J, and Ozen T 2003 Proc. of the Int. Conf. on Australian and New Zealand Intelligent Information Systems 233-238
[5] Tsuen-Ho H, Kao-Ming C, and Hei-Chun C 2000 Proc. of the Ninth IEEE Int. Conf. on Fuzzy Systems 621-626
[6] Ghosh S and Dubey S K 2013 Int. J. of Advanced Computer Science and Applications 4 36-39
[7] Savic D A and Walters G A 1999 Anti-Corrosion Methods and Materials 46 415-425
[8] Makropoulos C and Butler D 2004 J. of Hydroinformatics 6 3-18
[9] Naizhuo Z and Yunsheng Z 2010 Proc. of the 2010 4th Int. Conf. on Bioinformatics and Biomedical Engineering 1-4
[10] Altunkaynak A, Özger M, and Çakmakci M 2005 Water Resources Management 19 641-654
[11] Şen Z and Altunkaynak A 2009 Expert Systems with Applications 36 11745-11752
[12] Corona-Nakamura M A, Ruelas R, Ojeda-Magana B, and Andina D 2008 Proc. of the World Automation Congress 1-9
[13] Corona-Nakamura M A, Ruelas R, Ojeda-Magana B, and Carr Finch D W 2009 Proc. of the IEEE Int. Conf. on Systems, Man and Cybernetics 3751-3756
[14] Hammouda K and Karray F 2000 SYDE 625: Tools of intelligent systems design, course project, University of Waterloo Canada
[15] Jipkate B R and Gohokar V V 2012 Int. J. of Computational Engineering Research 2 737-739
[16] Bezdek J C, Pattern Recognition with Fuzzy Objective Function Algorithms: Kluwer Academic Publishers, 1981.
[17] Azar A T, El-Said S A, and Hassanien A E 2013 Computer Methods and Programs in Biomedicine 111 1-16
[18] Mohd Hanif H, Rasmani K A, and Mohamed Ramli N 2013 Proc. of the IJAS American-Canadian Int. Conf. on Academic Disciplines 2013