Water quality comprehensive evaluation method for large water distribution network based on clustering analysis
Kui Chang, Jin Liang Gao, Wen Yan Wu and Yi Xing Yuan

ABSTRACT
In order to evaluate water quality for a large water distribution network comprehensively, a two-stage classification method was used and the clustering methods, self-organizing map (SOM), K-means method and fuzzy c-mean (FCM), were represented. With these clustering methods, the pipes of a large real water distribution network were divided into some groups considering one or more water quality indicators synchronously. The water quality indicators of residual chlorine, water age, THMs, TAAs, TOC and BDOC are used in this paper. Residual chlorine and water age are two main water quality indicators. THMs and TAAs can represents the disinfection byproducts information. And TOC and BDOC are used to represents biological stability. According to the clustering results, the status of water quality of the water network was analysed. The results showed that the classification of SOM could express the comprehensive water quality in a water distribution network (WDN) directly and vividly by high-dimension water quality indicator projection to a low dimensional topology grid and that two-stage classification method has higher efficiency in comparison to the traditional clustering method. Water quality comprehensive evaluation was of significance for locating water quality monitoring, water network rehabilitation and expansion.

Key words | clustering analysis, fuzzy c-mean, K-means, self-organizing map, water distribution network, water quality evaluation

NOTATION

<table>
<thead>
<tr>
<th>SOM</th>
<th>Self-organizing map</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCM</td>
<td>Fuzzy c-mean</td>
</tr>
<tr>
<td>WDN</td>
<td>Water distribution network</td>
</tr>
<tr>
<td>DBPs</td>
<td>Disinfection byproducts</td>
</tr>
<tr>
<td>X</td>
<td>Input vector</td>
</tr>
<tr>
<td>M</td>
<td>The dimension of the input pattern X</td>
</tr>
<tr>
<td>Wj</td>
<td>The synaptic weight vector of jth neuron in the Kohonen layer</td>
</tr>
<tr>
<td>l</td>
<td>The total number of neurons in the Kohonen layer</td>
</tr>
<tr>
<td>i(X)</td>
<td>The neuron closest to the corresponding input pattern X</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>dij</td>
<td>The Euclidean distance between the winning neuron i and the neighboring neuron j</td>
</tr>
<tr>
<td>σ(t)</td>
<td>The effective width of topological neighborhood at time t</td>
</tr>
<tr>
<td>hi(t)</td>
<td>The topological neighborhood at time t</td>
</tr>
<tr>
<td>η(t)</td>
<td>The learning rate at time t</td>
</tr>
<tr>
<td>Wi(t+1)</td>
<td>The synaptic weight vector of neuron j at time t+1</td>
</tr>
<tr>
<td>j2(U,P)</td>
<td>Objective function</td>
</tr>
<tr>
<td>µik</td>
<td>The membership degree, which is interpreted as the posterior probabilities of selecting the i-th cluster given the data point xk</td>
</tr>
<tr>
<td>(dik)2</td>
<td>The squared inner-product distance norm</td>
</tr>
<tr>
<td>PC</td>
<td>Partition coefficient</td>
</tr>
<tr>
<td>CE</td>
<td>Classification entropy</td>
</tr>
<tr>
<td>c</td>
<td>Clustering value</td>
</tr>
<tr>
<td>µij</td>
<td>The membership of data point j in cluster i</td>
</tr>
</tbody>
</table>

The number of observations $N$

The water quality indicator vector composed of residual chlorine, THMs, TOC, BDOC, water age, TAAs, UV-254, flow rate, velocity, and unit headloss $I$

The water quality indicator calculation value at $t$th hour $i_t$

THMs Trihalomethanes

TOC Total Organic Carbon

BDOC Biodegradable dissolved organic carbon

TAAs Total amino acids

UV-254 Ultraviolet light absorbance at 254 nm

The number of the pipes of the water supply network $n$

The average quality error $q_e$

Topology error $t_e$

The learning rate at time $t$ $\eta(t)$

The davies-bouldle index DBI

The distance between the element within the group $A$

The distance between the groups $B$

The weighting exponent which determines the fuzziness of the clusters $m$

The termination tolerance of the clustering method $\epsilon$

INTRODUCTION

With economic development and reinforcement of sanitation consciousness, attention has been drawn to water quality security. A water distribution network (WDN) which connects water plant and users is important infrastructure. The quality of water in a WDN is of significance to supply good drinking water. A large number of water quality indicators which are based on three international authorized drinking water standards are selected to ensure the water quality. But it is impossible to monitor or measure all the water quality indicators for the whole WDN. What is more, a single water quality indicator is rather one-sided. So it is necessary to make comprehensive evaluation on water quality of WDN.

A comprehensive water quality evaluation system is proposed in this paper. It is of great significance to guarantee the safety of water quality in a water supply network. Models for DBPs (disinfection byproducts) have been developed for different purposes (Sadiq & Rodriguez 2004a, b). Hydroinformatics tools are applied for water quality modeling and management by Mark et al. (2010). With the development of a WDN water quality model, the WDN water quality evaluation system has made great progress. Horton published the first indicator of water quality evaluation systems and proposed water quality evaluation methods in 1965 (Horton 1965). Principal component analysis, factor analysis and discrimination analysis were used to evaluate spatial and temporal variations in water quality (Palani et al. 2008). Several artificial intelligence methods, such as neural network, genetic algorithm, fuzzy system, support vector machines, and data mining, have been applied in water resource engineering (Cheng et al. 2002; Lin et al. 2006; Muttil & Chau 2006; Chau & Muttil 2007; Wang et al. 2009; Wu et al. 2009).

Some researchers have investigated complex system evaluation methods and derived some preliminary results in China. A series of complex system evaluation methods are proposed, such as the grey system theory, and matter-element analysis theory. The values of the evaluation index for real water supply networks are generally continuous real numbers. It is difficult to express the relationship between the indices, and the evaluation results may be incompatible. The water quality level is assessed according to the pessimistic principle in relation to the existing national water quality standards in China: the comprehensive water quality category is determined by the worst single indicator. Such evaluation is too extreme and one-sided. The results are greatly affected by subjective factors and lack of objectivity. In this paper, self-organizing feature map or self-organizing map (SOM) neural network, K-means and the FCM clustering algorithm are used for urban WDN cluster analysis, and a comprehensive WDN water quality evaluation system is built.

CLUSTERING METHODS

Self-organizing feature map

The structure of self-organizing feature map

The SOM network, which was first proposed by Kohonen in 1981 (Kohonen 1981), is a non-supervised learning neural
network model, followed by a wide range of applications (Kohonen 1982, 1989, 1995). The structure of the SOM is shown in Figure 1. The SOM network is composed of two layers. The top one is the input layer and the bottom one is the output. The number of neurons in the input layer is equal to the number of variables it contains, and its function is to obtain data. The output neurons constitute a one-dimensional or two-dimensional grid. In the figure, the output neurons are in a two-dimensional grid space. The relationships between the neurons in the space are identified by the grid.

The concept of the learning algorithm used for training the SOM is unsupervised and competitive. The SOM training process is briefly described below.

The input pattern of the SOM are denoted by

\[ X = [x_1, x_2, \ldots, x_M]^T. \] (1)

The synaptic weight vectors of neurons of the Kohonen layer are denoted by

\[ W_i = [w_{i1}, w_{i2}, \ldots, w_{iM}]^T, \quad j = 1, 2, \ldots, l. \] (2)

The synaptic weights are initialized as small random numbers at the beginning of the training process. The competitive learning signifies that neurons within the network compete with each other to determine which one will be activated to become the winning neuron. The magnitudes of the similarities between an input pattern and the synaptic weights of all neurons are the indices used to determine the winning neuron. Thus, the determination of the winning neuron can be accomplished using (Kohonen et al. 2001)

\[ i(X) = \arg \min_j \|X - W_j\|. \quad j = 1, 2, \ldots, l \] (3)

\[ \|X - W_i\| = \sqrt{\sum_{i=1}^{M} (x_i - w_{ji})^2}, \quad i = 1, 2, \ldots, l. \] (4)

The Euclidean distance is often used as the similarity measure for SOM. Smaller \( \|X - W_i\| \) means higher similarity between the input pattern \( x \) and the synaptic weight vector \( W_i \). Therefore, the strategy used to determine the winning neuron is to calculate the distances from the current input pattern to all output neurons of the Kohonen layer and find the smallest distance. The corresponding output neuron with the smallest distance to the current input pattern is the winning neuron.

After determining the winning neuron, the lateral interactions between the winning neuron and its neighborhood are then considered. The influence of the winning neuron on its neighboring neurons is calculated using the topological neighborhood function (Juha et al. 1999). A representative topological neighborhood function is

\[ h_{j,t}(X) = \exp \left( -\frac{d_{ji}^2}{2\sigma^2(t)} \right), \quad t = 0, 1, 2, \ldots \] (5)

Next, the values of the synaptic weights are adjusted according to input patterns using the algorithm defined as

\[ W_j(t + 1) = W_j(t) + \eta(t)h_{j,t}(X)(X - W_j(t)). \] (6)

The learning rate will shrink with time. By repeatedly performing the above process, the winning neuron and the neighboring neurons become more similar to the corresponding input pattern. Finally, a trained SOM can be obtained. More details regarding the learning algorithm of SOM are given by Kohonen et al. (2001).

Each input neuron node is connected with all the output neurons. When the Euclidean distance between the input and the weight vector becomes minimized, the right output neuron is activated as the network output. Then the weight vector is amended to be closer to the input vector. The output neuron is called the winning neuron. The weights of the corresponding area of the right output neurons are also amended until the network gets the terminating conditions.

Figure 1 | Model of self-organizing feature map.
The features of the self-organizing feature map

Compared with other classification methods, SOM is a self-organizing and self-learning network, and it can be used for real-time study. The network has stability and strong anti-noise capacity, without the evaluation function, and can identify the most significant features of vector space (Juha & Esa 2000). As a competition-based neural network, the main characteristics of SOM neural networks are (1) the nature of self-organization, that is the ability to maintain topology; (2) the nature of self-organizing probability; it can form the probability distribution density of the neurons which corresponds to the spatial distribution density of the samples.

The application of K-means algorithm in self-organizing feature map

The K-means clustering algorithm is used after the process of the SOM neural network. The training sample set is classified by the SOM neural network, and the corresponding adjustments to the weight value are taken. Then the representative neurons are found. The weight values of these neurons are classified by the K-means clustering algorithm. The eigenvector set for one category is at the basis of the weight values of neurons of competition layer. So a class is expressed by a group of representative eigenvector. Therefore, it integrates the advantages of the SOM network and K-means clustering algorithm and achieves better results.

Fuzzy c-mean

Fuzzy c-means (FCM) was introduced by Dunn (1974) and generalized by Bezdek (1981). The performance index (objective function) guiding the search through the data space assumes the form

$$ J_2(U, P) = \sum_{i=1}^{C} \sum_{k=1}^{n} \mu_{ik}^2 d_{ik}^2, \text{ s.t. } U \in E_i. $$

Cluster validity refers to the problem of whether a given fuzzy partition fits to all the data. The clustering algorithm always tries to find the best fit for fixed number of clusters and the parameterized cluster shapes. Different scalar validity measures have been proposed in the literature, but none of them is perfect on its own. Partition coefficient (PC) and classification entropy (CE) are used in this paper.

**Partition coefficient**

PC measures the amount of “overlapping” between clusters. It is defined by Bezdek as follows:

$$ PC(c) = \frac{1}{N} \sum_{i=1}^{C} \sum_{j=1}^{N} \mu_{ij}^4. $$

The disadvantage of PC is the lack of direct connection to some property of the data themselves. The optimal number of clusters is at the maximum value.

**Classification entropy**

CE measures the fuzziness of the cluster partition only, similar to the PC:

$$ CE(c) = \frac{1}{N} \sum_{i=1}^{C} \sum_{j=1}^{N} \mu_{ij} \log(\mu_{ij}). $$

Selection of water quality evaluation index in water distribution network

To construct a scientific and rational clustering evaluation system for the water supply network, some water quality indicators which can reflect the water quality in the pipe are selected for the clustering evaluation. The chemical and the microbiological risk are two major risks for water supply network. Chemical risks are mainly produced by the disinfectants and disinfection by-products. Microbiological risk is due to bacteria regeneration in the pipeline. The relevant amendments of water quality standards in China and other countries also fully embody such change.

Research shows that there is some interrelated relationship between the disinfectants, disinfection by-products and organic matter and bacteria. The main indicators of residual chlorine, THMs, TOC, BDOC, water age, TAAs, UV-254, flow rate, velocity are easy to long-term monitor. What is more, unit headloss is also used to consider WDN physical status. To a certain extent, these indicators can represent the
whole WDN water quality. Compared with other indicators, they are easy to measure and simulate.

In order to get hydraulic and water quality information at each pipe, unit headloss, flow rate, velocity, water age and residual chlorine were calculated by EPANET (Rossman 1994) and THMs, TOC, BDOC, TAAs and UV-254 were calculated by related formulas on the basis of water age and residual chlorine value. The value of hierarchical water quality indicators are selected every one hour, because these indicators make fluctuations during the 24 hours of a day. So the input pattern of SOM was

\[ X = [I_1, I_2, ..., I_i, ..., I_{10}] \]

\[ I = [i_1, i_2, ..., i_1, ..., i_{24}] . \]

If the number of the pipes of the water supply network is \( n \), the number of pipes of a better topology network is \( 5 \times n^{1/2} \). But there are many such topology networks. And ultimate network structure is determined by \( qe \) and \( te \), which are received by the temporary training of typical structure. And the topology nodes are with a rectangular form. Learning rate is calculated using the following formula:

\[ \eta(t) = 0.9(1 - t/1000) . \]  

There are 5449 pipes on the water distribution network, so the number of neurons of a better network topology is

**ESTABLISHMENT OF WATER QUALITY CLUSTERING EVALUATION SYSTEM FOR WATER DISTRIBUTION NETWORKS**

**Water distribution network introduction**

Figure 2 shows the large water distribution network, which is used for the analysis of water supply network, selected for the water quality clustering evaluation analysis in this paper. The water distribution network is composed of 4542 nodes, 7 reservoirs, 5449 pipes, and 48 pumps.

The input vector of SOM algorithm is composed of the data of water quality and hydraulic indicators. The water quality indicators are the residual chlorine, water age, THMs, TAAs, TOC and BDOC, and the hydraulic indicators are demand and pressure. Residual chlorine and water age are two main water quality indicators. THMs and TAAs can represent the disinfection byproducts information. TOC and BDOC are used to represents biological stability.

**Water quality evaluation based on SOM**

**The building of self-organizing feature map**

The water quality clustering analysis of water supply network is achieved with Matlab toolbox and the results are expressed in a visual way. The different values of the parameters are used in the SOM model of water supply network in order to get a better training result.

**Figure 2** Water distribution network topology.
about 370. Learning rate is 0.85. Initial weight value is randomly assigned. After the temporary training of typical structure, $qe$ and $te$, are shown in Table 1. So the default number of neurons is 384.

The Gaussian function is taken as the winning neighborhood function, and neighbor areas reduce with a square form. SOM algorithm takes the way of batch training, which is divided into two stages: the rough and the fine tuning stage. Each sample is usually trained at least 20 times in the network training.

**The training of self-organizing feature map and clustering process**

The trained network topology represents the final state of pipes of the water supply network after the process of learning and training. According to the weight value of the trained network, the distance between the adjacent neurons is calculated. As shown in Figure 3, the U-matrix (Ultsch & Siemon 1990) of the SOM network is received. The plans of the variable for each indicator are received by reversing the standardization process of the variable. Each neuron in the figure contains more than one pipe in which the water quality is similar.

**The results evaluation**

The flow chart of water quality clustering evaluation system is shown in Figure 4. The water supply network is divided into different types of the region using the U-matrix, and then cluster analysis is taken with the K-means algorithm. The number of clustering is automatically selected by DBI. DBI is calculated using the following formula:

$$DBI = \frac{\sum A}{B}.$$  \hspace{1cm}(13)

So the DBI is at the minimum for the best clustering results. The number of class is selected and the result is expressed in a visual way.

According to the weight value of the trained SOM neurons, the state of water quality in the pipes is classified with the K-means and SOM algorithm, and the DBI is selected as the comprehensive indicator for K-means method, and PC and CE for FCM.

**The analysis of the results**

The input vector is processed by the SOM algorithm, and the plans for the variable of the indicators are achieved. Each

<table>
<thead>
<tr>
<th>Network</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$qe$</td>
</tr>
<tr>
<td>$19 \times 19$</td>
<td>0.395</td>
</tr>
<tr>
<td>$20 \times 18$</td>
<td>0.395</td>
</tr>
<tr>
<td>$22 \times 17$</td>
<td>0.392</td>
</tr>
<tr>
<td>$24 \times 16$</td>
<td>0.386</td>
</tr>
<tr>
<td>$26 \times 15$</td>
<td>0.410</td>
</tr>
<tr>
<td>$28 \times 14$</td>
<td>0.405</td>
</tr>
<tr>
<td>$30 \times 13$</td>
<td>0.405</td>
</tr>
<tr>
<td>$32 \times 12$</td>
<td>0.409</td>
</tr>
<tr>
<td>$34 \times 11$</td>
<td>0.408</td>
</tr>
<tr>
<td>$36 \times 10$</td>
<td>0.409</td>
</tr>
</tbody>
</table>

Figure 3 | U-matrix
neuron on the plans contains more than one pipes in which the water quality is similar. The component plans are shown in Figure 5.

As shown in the U-matrix and component plans map, all the neurons are divided into five classes. For example, the neurons located in the lower left corner of the component plans are with smaller concentrations of chlorine, and larger TAAs, water age, THMs, TOC, UV254 and BDOC values. And the result is just opposite at the top of the component plans. The mean value of each category is shown in Table 2.

To verify the results of these intuitive judgments, K-means and FCM cluster methods are used for the trained neurons.

The water quality in the pipes is classified according to the comprehensive indicator of DBI. The result is shown in Figures 6–9. The result shows that the DBI value is smallest at the number of 5. As shown in Figure 8, the cluster result and the intuitive judgments are consistent.

Water quality evaluation based on FCM

The water quality is also evaluated by the FCM algorithm. The optimization parameters were fixed to the following values: \( m = 2, \epsilon = 0.0001 \) for each cluster, \( c \in [2, 20] \). The values of the validity measures depending from the number of cluster are plotted in Figure 10.

We must mention that no validation index is reliable only by itself, so that is why all the PC and CE indexes are shown, and the optimum can be only detected with the comparison of all the results. We consider that partitions with fewer clusters are better, when the differences between the values of a validation index are minor.

The main drawback of PC is the monotonic decreasing with \( c \) and the lack of direct connection to the data. CE has the same problems: monotonic increasing with and hardly detectable connection to the data structure. On the score of Figure 11, the number of clusters can be only rated to 4.

CONCLUSION

As water utilities evolve from having the single mission of supplying high quality water to consumers and also ensuring water security, water network analysis will also need to
evolve to better simulate and evaluate the water quality. Traditional evaluation methods determine the quality of drinking water according to a single water quality indicator. It is difficult to know about the comprehensive status of water quality in WDN. In order to better understand the comprehensive water quality status, the water quality comprehensive evaluation method for large water distribution network based on clustering analysis is presented in this paper.

The water quality indicators of residual chlorine, THMs, TOC, BDOC, water age, TAA, UV-254, flow rate, velocity, and unit headloss were selected to comprehensively evaluate water quality of WDN. Unit headloss, flow rate, velocity, water age and residual chlorine were calculated by EPANET. Then THMs, TOC, BDOC, TAA and UV-254 were calculated by those results. A two-stage classification method was used in this paper. All these results were processed by SOM

<table>
<thead>
<tr>
<th>Table 2</th>
<th>The water quality index mean value of each category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
<td>1</td>
</tr>
<tr>
<td>Chlorine (mg/L)</td>
<td>0.91</td>
</tr>
<tr>
<td>Water age (hr)</td>
<td>2.36</td>
</tr>
<tr>
<td>THM (mg/L)</td>
<td>4.87</td>
</tr>
<tr>
<td>TAA (µg/L)</td>
<td>2.26</td>
</tr>
<tr>
<td>TOC (mg/L)</td>
<td>3.37</td>
</tr>
<tr>
<td>UV-254 (cm-1)</td>
<td>0.071</td>
</tr>
<tr>
<td>BDOC (mg/L)</td>
<td>0.23</td>
</tr>
</tbody>
</table>
firstly, and the component plan for each water quality indicator was gained. The topology network of all the pipeline was also be built. By the component plans, we can get a general overview of the situation of water quality in WDN. By the SOM results, two methods, K-means and FCM, were respectively used to get a better clustering way of all the pipelines in WDN.

The result shows that water quality comprehensive evaluation is a better method to assess the water quality distribution in a WDN, and that the SOM method can express the comprehensive water quality in WDN directly and vividly by high-dimensional water quality indicator projection to a low-dimensional topology grid. Compared with the traditional clustering method, this two-stage classification method has higher efficiency and both K-means and FCM methods gained a better water quality clustering result.

SOM is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional, discretized representation of the input space of the training
samples. The relation between the convergence of the SOM algorithm and the learning rate and weight values, and how to determine the optimal learning rate needs further study.

ACKNOWLEDGEMENTS

The work described in this paper was funded by the China National Science and Technology Support Programme under Grant 2006BA08B03, Harbin Institute of Technology under Grant: HIT.NSRIF.2008.68, and a UK–China science network grant from the Royal Society, London.

REFERENCES


First received 25 January 2010; accepted in revised form 23 September 2010. Available online 3 February 2011