Stream Discharge Analysis Model Based on Fuzzy Clustering Iterative Method

Rong He
School of Civil Engineering and Communication, North China University of Water Resources and Electric Power, Zhengzhou 450045, Henan, China

Abstract
In order to rationally develop rivers, it is necessary to observe stream discharge and predict it based on a large amount of measured data. Data mining technology can be used to obtain useful information from the measured data. But the computational efficiency decreases with the increasing of quantities of data. To improve computational efficiency and accuracy, a new method is proposed in this paper which combines the fuzzy C means clustering algorithm and the fuzzy clustering cyclic iterative technique to construct a fuzzy clustering iterative model (FCIM). The model is used to cluster, identify and analyze the observation data. In the calculation, the sample set is firstly normalized to determine the membership matrix, class membership matrix and clustering center. Then according to characteristic values, number of clusters and accuracy of calculation cluster the data, give the weight vector and perform iterative calculation. A river is taken as a case to verify the analysis model. The result shows that the method can effectively and accurately reveal the type and the rule of stream discharge, and predict the river flow which provides an important reference for rationally development of rivers.

Key words: Stream Discharge, Data Mining, Clustering Algorithm

1. INTRODUCTION
Nowadays, the general method of large-scale data processing can only be used to query and statistics, and can't analyze the intrinsic relation and implied information of the data. Data mining technology can intelligently and automatically obtain available information and knowledge from the data (Gong, Tian and Chen, 2010; Wang, Si, Yang, 2005; Pahl, 2004). But when the amount of data is becoming larger and larger, the computational efficiency decreases which also affect the analysis and decision-making.

Clustering is an important data analysis technology which can divide a collection of abstract objects into numerous classes of similar objects. Clustering analysis has been widely used in data analysis, pattern recognition, image processing, market research and so on (Everitt and Hothorn, 2009; Diday and Simon, 1976; Jung, Kim and Sim, 2016). By clustering, the area of dense or sparse is identified firstly, and then the global distribution pattern and the relationship between the data attributes can be gained (Moulton, Roberts, Calamai, 2009).

Fuzzy set theory can be used to do fuzzy evaluation, fuzzy decision, fuzzy pattern recognition and fuzzy analysis in the practical problems. Generally the higher the complexity of the system is, the stronger the ambiguity is. Fuzzy set theory is used in the membership degree to describe the indistinguishable character of the fuzzy things. In 1973, Dunn put forward Fuzzy C mean clustering algorithm (FCM) (Dunn, 1973). Bezdek J.C. and other scientists developed it and used the method extensively in the practical application (Bezdek, 1981; Höppner, Klawonn and Kruse, 2000). The fuzzy clustering iterative model (FC) created by Chen S.Y. is widely used in the field of water resources (Chen, 2004). Fuzzy clustering analysis (FCA) is based on the characteristics of the object to study the object classification and predict the samples. The principle is to make the objects of the same class be more similar and the different objects be more different. The method of fuzzy clustering is to solve the problem and meet the requirements of the aggregation in the classes. Before clustering analysis, these categories are potential and the number of clusters is unknown.

In China, there is lack of relevant data of dam break and the data has poor uniform continuity at present. The variable fuzzy set theory is introduced to predict life loss of dam break. And the population mortality prediction model is established by using variable fuzzy clustering iterative method. Then adjacent fuzzy scale method is adopted to determine the index weight. At the same time optimal relative membership degree matrix and optimal clustering center matrix are obtained with the model. But the model has only a certain practical and accuracy (Liu and Li, 2011).

The emergence of fuzzy time series has recently received more attention because of its capability of dealing with vagueness and incompleteness inherent in data. But deriving an effective and useful forecasting model is a challenge task. In the previous work, some researchers put forward two crucial issues, namely controlling uncertainty and effectively partitioning intervals, as well as develop a deterministic forecasting model to manage these issues (Li, Kuo and Cheng, 2011).
To improve computational efficiency and accuracy, a new method is proposed in this paper which combines the fuzzy C means clustering algorithm and the fuzzy clustering cyclic iterative technique to construct a fuzzy clustering iterative model (FCIM). This paper proposes the fuzzy clustering iterative model (FCIM), and gives a simple and convenient method to determine the optimal number of clusters which is applied to predict the stream discharge of a river.

2. FUZZY CLUSTERING ITERATIVE MODEL

2.1 Fuzzy clustering iterative algorithm

For the clustering sample set consisting of \( n \) samples, the order eigen value matrix can be expressed as:

\[
Y_{max} = \begin{bmatrix}
Y_{11} & Y_{12} & \cdots & Y_{1n} \\
Y_{21} & Y_{22} & \cdots & Y_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
Y_{m1} & Y_{m2} & \cdots & Y_{mn}
\end{bmatrix} = \begin{bmatrix} y_{1y} \\
y_{2y} \\
\vdots \\
y_{ny}
\end{bmatrix} \tag{1}
\]

where \( y_{iy} \) is the characteristic value of index \( i \) for the sample \( j \); \( i = 1,2,\ldots,m; \) \( j = 1,2,\ldots,n \).

To eliminate the dimension difference of index \( M \) before clustering, the characteristic value of the index can be transformed into the relative membership degree to \( \tilde{A} \). According to the property of problem the commonly normalized formulas include the bigger optimizing target and the smaller optimizing target. The index matrix can be changed into the corresponding membership matrix as:

\[
\beta_{max} = \begin{bmatrix}
\beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\
\beta_{21} & \beta_{22} & \cdots & \beta_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
\beta_{n1} & \beta_{n2} & \cdots & \beta_{nm}
\end{bmatrix} = \begin{bmatrix} \beta_{1y} \\
\beta_{2y} \\
\vdots \\
\beta_{ny}
\end{bmatrix} \tag{2}
\]

where \( \beta_{iy} \) is relative membership coefficient.

Supposed according to \( c \) categories and \( m \) index characteristic values \( n \) samples to be clustered, the type membership matrix is:

\[
X_{cmax} = \begin{bmatrix}
x_{11} & x_{12} & \cdots & x_{1c} \\
x_{21} & x_{22} & \cdots & x_{2c} \\
\vdots & \vdots & \ddots & \vdots \\
x_{c1} & x_{c2} & \cdots & x_{cn}
\end{bmatrix} = \begin{bmatrix} x_{1y} \\
x_{2y} \\
\vdots \\
x_{cy}
\end{bmatrix} \tag{3}
\]

The constraints condition is expressed as:

\[
\sum_{h=1}^{c} x_{hy} - 1 = 0, \quad 1 > x_{hy} > 0, \forall j, \quad h=1,2,\ldots,c
\]

where \( x_{hy} \) is the relative membership degree of sample \( j \) to category \( h \). The cluster centers of \( C \) classes could be expressed as:

\[
D_{max} = \begin{bmatrix}
d_{11} & d_{12} & \cdots & d_{1c} \\
d_{21} & d_{22} & \cdots & d_{2c} \\
\vdots & \vdots & \ddots & \vdots \\
d_{c1} & d_{c2} & \cdots & d_{cn}
\end{bmatrix} = \begin{bmatrix} d_{hy} \\
d_{2y} \\
\vdots \\
d_{cy}
\end{bmatrix} \tag{4}
\]

Then we could define the fuzzy loop iterative clustering model as follows:

\[
x_{hy} = \begin{cases} 
0 & (\Delta_{hy} = 0, k \neq h) \\
\frac{\sum_{i=1}^{n}\alpha_{i}(\beta_{iy} - d_{ky})^2}{\sum_{i=1}^{n}\alpha_{i}(\beta_{iy} - d_{ky})^2} & (\Delta_{hy} \neq 0) \\
1 & (\Delta_{hy} = 0)
\end{cases} \tag{5}
\]
Given weight vector, then we could compose the fuzzy clustering iterative model based on Eqs. (4) and (5). The steps of solving the fuzzy clustering iterative model are as follows:
Firstly, give clustering number \( c \) and iterative computation precision \( \delta_1, \delta_2 \);
Secondly, set initial fuzzy clustering matrix \( x_{ij}^0 \) and initial fuzzy clustering center matrix \( d_{ih}^0, l = 0 \);
Thirdly, use the above model to calculate \( x_{ij}^{l+1}, d_{ih}^{l+1} \);
Fourthly, if \( \max | x_{ij}^{l+1} - x_{ij}^l | \leq \delta_1, \max | d_{ih}^{l+1} - d_{ih}^l | \leq \delta_2 \), then the iterative end; \( x_{ij}^{l+1}, d_{ih}^{l+1} \) should be used as the optimal fuzzy clustering center matrix \( x_{ij}^* \) and the optimal fuzzy clustering matrix \( d_{ih}^* \) which can satisfy the calculation accuracy \( \delta_1 \) and \( \delta_2 \). Otherwise, continue to go on the iterative calculation following the third step.
Lastly, when the accuracy requirements are meet, the iterative is end.

### 2.2 Determination of Optimal Cluster Number
The iterative of fuzzy clustering model is always under the condition that the clustering number \( C \) is known. When the class number is unknown, it is necessary to determine the optimal number \( C \) of clusters. Supposed the number of categories is gradually increasing, it is obviously that the criterion function \( F(u, s, w) \) is gradually decreasing with the increase of the number of classes. When the sample set has \( C \) concentrated classes (the optimal number of clusters), the criterion function \( F(u, s, w) \) decreases rapidly from one class to the \( C \) class; When the number of clusters continue to increase, that is, which means the original dense group separating repeatedly. At this time, the criterion function continues to decrease, but the speed is slow until \( c=N \) (sample number) and then the criterion function is the smallest which means \( F(u, s, w) =0 \). So the number of categories corresponding to the inflection point can be considered as the optimal number of clusters \( C \).
In the application, the numbers of clusters \( C \) from 1 to \( N \) are calculated by the method of using the fuzzy loop iteration (\( N \) as an integer, selected according to the number of samples, generally take ranged from 15 to 20). Draw the curve of function \( F(u, s, w) \) —\( C \), then judge the break point and the class corresponding to the inflection point is the best clusters number.

### 2.3 Sample Classification
According to the inapplicability of the maximum membership principle under classification condition (2), the class can be determined by the method of level character value. The level character value can be got by Eq.(6).

\[
T = (1, 2, 3, \ldots, c)^* = \begin{bmatrix}
    t_{11} & t_{12} & \cdots & t_{1c} \\
    t_{21} & t_{22} & \cdots & t_{2c} \\
    \vdots & \vdots & \ddots & \vdots \\
    t_{c1} & t_{c2} & \cdots & t_{cc}
\end{bmatrix} = (T_1, T_2, \ldots, T_c)
\]

Vector \( T \) is the level character value. According to the level character value the sample can be classified as:

\[
y_i \in C \quad c - 0.5 < T_i \leq c + 0.5
\]

### 3. APPLICATION OF FCIM IN THE RESEARCH OF RIVER DISCHARGE
A river located in China. Recently the exploiting of rivers is an important task of water resources development, so it is necessary to research the discharge regularity of the river. River discharge is impacted by many factors, such as climate conditions, underlying factors and so on. The method of FCIM mentioned in this paper is used to analyze it.
The clustering analysis is taken on 4 relevant factors corresponding to the measured flow of the river during 23 years. Factor 1 is the weather station-total rainfall from November of previous year to March of current year; Factor 2 is the monthly mean zonal lateral circulation index in the Eurasian region of the last August; Factor 3 is the monthly mean zonal radial circulation index in the Eurasian region of the last May; Factor 4 is the 2800MHz solar radio flux of last June.

After correlation analysis, factors 1, 3, 4 are direct proportion with the measured runoff, which belongs to the type of the bigger the better; factor 2 is negative correlated with the measured runoff, which belongs to the type of the smaller the better. All factors are formatted by the appropriate formula.

To determine the weights of the four indicators, the method of binary comparison fuzzy decision analysis with the factor correlation are used to determine the weight of the index. And the four index relative weight vectors are as follows:

\[ \alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.345, 0.32, 0.21, 0.16) \]  

(8)

To maintain unchanged, S and U are iterated cyclically according to Eq.(5). The convergence accuracy is $10^{-5}$, and clustering number is from 1 to 12.

Then analyses the optimal cluster numbers, calculates $F = F(u, s, w)$, the relations are showed in table (1):

### Table 1. F–C relations

<table>
<thead>
<tr>
<th>Clustering number C</th>
<th>$F$</th>
<th>Clustering number C</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1151</td>
<td>7</td>
<td>0.0045</td>
</tr>
<tr>
<td>2</td>
<td>0.0366</td>
<td>8</td>
<td>0.0042</td>
</tr>
<tr>
<td>3</td>
<td>0.0231</td>
<td>9</td>
<td>0.0044</td>
</tr>
<tr>
<td>4</td>
<td>0.0111</td>
<td>10</td>
<td>0.0034</td>
</tr>
<tr>
<td>5</td>
<td>0.0072</td>
<td>11</td>
<td>0.0031</td>
</tr>
<tr>
<td>6</td>
<td>0.0051</td>
<td>12</td>
<td>0.0023</td>
</tr>
</tbody>
</table>


![Figure 1](image-url)  

Figure 1. The relation curve of the $F_j(u, s, w)$ - C in runoff clustering

When $C=4$, the clustering center is obtained by clustering iterative computation:

\[
D = \begin{bmatrix}
0.9482 & 0.5245 & 0.4597 & 0.2968 \\
0.7359 & 0.7912 & 0.6209 & 0.6102 \\
0.7156 & 0.8402 & 0.8257 & 0.6567 \\
0.5600 & 0.8541 & 0.3568 & 0.4342
\end{bmatrix}
\]

(9)

We can calculate the clustering level character values of runoff samples in 23 years respectively which are showed in Table (2) and classify them.

According to Table 2 and Figure 2, when the class number of the river flow sample is 4, combining the clustering level character values with the data of the cluster center, we can classify the type of the river discharge by Eq. (7) as follows: Sample \{17\} belongs to Class 1; Sample \{4,6,7,8\} belongs to Class 2, Sample \{1,2,3,5,11,12,16,18,19,20,21\} belongs to Class 3, Sample \{9,10,13,14,15,22,23\} belongs to Class 4. So the type of the river discharge is mainly composed of Class 3 and 4, less composed of Class 2 and 1.
Table 2. The classified runoff of the river

<table>
<thead>
<tr>
<th>Sample</th>
<th>Level value</th>
<th>Sample</th>
<th>Level value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.989</td>
<td>13</td>
<td>3.679</td>
</tr>
<tr>
<td>2</td>
<td>2.982</td>
<td>14</td>
<td>3.679</td>
</tr>
<tr>
<td>3</td>
<td>3.042</td>
<td>15</td>
<td>3.829</td>
</tr>
<tr>
<td>4</td>
<td>2.247</td>
<td>16</td>
<td>3.461</td>
</tr>
<tr>
<td>5</td>
<td>3.4021</td>
<td>17</td>
<td>1.040</td>
</tr>
<tr>
<td>6</td>
<td>2.169</td>
<td>18</td>
<td>2.852</td>
</tr>
<tr>
<td>7</td>
<td>2.1771</td>
<td>19</td>
<td>3.332</td>
</tr>
<tr>
<td>8</td>
<td>2.046</td>
<td>20</td>
<td>2.791</td>
</tr>
<tr>
<td>9</td>
<td>3.561</td>
<td>21</td>
<td>2.759</td>
</tr>
<tr>
<td>10</td>
<td>3.648</td>
<td>22</td>
<td>3.643</td>
</tr>
<tr>
<td>11</td>
<td>3.489</td>
<td>23</td>
<td>3.669</td>
</tr>
<tr>
<td>12</td>
<td>3.089</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. The curve of runoff of the river in 23 years

4. CONCLUSION

In this paper, a new method is proposed which combines the fuzzy C means clustering algorithm and the fuzzy clustering cyclic iterative technique to construct a fuzzy clustering iterative model (FCIM). The model is used to cluster, identify and analyze the observation data. In the calculation, the sample set is firstly normalized to determine the membership matrix, class membership matrix and clustering center. Then according to characteristic values, number of clusters and accuracy of calculation cluster the data, give the weight vector and perform iterative calculation. A river is taken as a case to verify the analysis model. The result shows that the method can effectively and accurately reveal the type and the rule of stream discharge, and predict the river flow which provides an important reference for rationally development of rivers.

The FCIM and the determination of optimal number of clusters are introduced in this paper, and the model is applied to research the regularity of river discharge which generally indicate the types and regularity of river flow. It also manifests that the FCIM is practicable and provides an important reference for further rational development of a river.

Acknowledgements

This work is supported by the Basic and Advanced Technology Research Project of Henan Province, China (No.152300410241).
REFERENCES