A new fuzzy c-means clustering-based time series segmentation approach and its application on tunnel boring machine analysis

Xueguan Song a, Maolin Shi a,*, Jianguo Wu b, Wei Sun a

a School of Mechanical Engineering, Dalian University of Technology, Linggong Road, Dalian 116024, China
b College of Engineering, Peking University, No.60 Yannan Yuan, Bejing 100871, China

ARTICLE INFO

Article history:
Received 13 August 2018
Received in revised form 24 May 2019
Accepted 25 July 2019

Keywords:
Tunnel boring machine
Time series segmentation
Fuzzy c-means clustering
Prior information

ABSTRACT

Tunnel boring machine (TBM) is a complex engineering system widely used for tunnel construction. In recent years, massive in-situ time series data of TBM has been recorded, which can provide important references and useful information for TBM designers and operators. In this work, a new fuzzy c-means clustering-based time series segmentation approach is proposed for TBM time series data, where the prior information of attributes is incorporated to facilitate effective segmentation. In this approach, the segmentation objective function is formed by multiplying the time distance and the spatial distance between data. The prior information, i.e. the torque of cutterhead, is correlated with the penetration rate, is described by a linear model and included in the part of spatial distance between data. A new decision making method based on the distance between the joint segment prototypes is proposed to determine the appropriate number of segments. The application on TBM time series data from a tunnel in China shows that the proposed approach can accurately identify different excavation status of the TBM, and help the other data mining tasks of TBM as well. The proposed approach also has promising applications to other complex engineering systems.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

Tunnel boring machines (TBMs) have been widely used for tunnel construction because of their relatively high efficiency, safety, and environmental friendliness compared with conventional blasting excavation [1]. A TBM is a large mechanical system which consists of several sub-systems, e.g., cutterhead, chamber and screw conveyor, and often works under various harsh environments [2]. Due to its complex interaction with the geology and various operation states, it is usually impractical to use physical-mathematical models to analyze and predict the behavior of TBM during the real excavation process [3–6]. With the recent advancement of cyber-physical systems and sensing technologies, now massive amounts of time series data can be readily obtained during the excavation, which provides unprecedented opportunities for data-driven design, analysis, and control of TBM [5–10]. To fulfill this aim, the very first and critical step is to partition the time series data into different parts based on data patterns or characteristics, such that each part corresponds to a certain TBM operation state. The task of partitioning a time series data is often referred to as time series segmentation (TSS) [11–14]. TSS is a process of segmenting a given time series to detect homogeneous segments with similar statistical characteristics and widely used in different areas [11–15]. Higgs collected the time series data of vehicle sensors and used TSS to partition them to recognize
the driving behaviors and account for the effects of unknown driving state factors [12], Jamali applied time series segmentation to analyze the vegetation time series data to help global climate change studies [13]. Burrell used time series segmentation to detect the change points of dryland degradation, and the results indicated that TSS is a useful tool for the studies of ecosystem change [14]. Wang used time series segmentation to detect the temporal and spatial trends of annual and seasonal land surface temperature data [15]. Hallac used temporal consistency to represent the sequential relationship of data and proposed a new time series segmentation approach [16].

In recent years, the data clustering-based TSS approaches have been widely used in the field of engineering since its good segmentation performance for high-dimensional time series data [17,18]. Fuzzy c-means algorithm (FCM) is a popular clustering approach and has been successfully used in various engineering areas [19–23]. Rosen used FCM to classify the operation states of a biological treatment process [19]. Seera used FCM to assist in power quality monitoring [20]. Li introduced FCM to the risk analysis of a dam [21], and Chen used FCM for partial discharge detection to find high-voltage cable defects [22]. Liu modified FCM and applied it to fluid identification in carbonate reservoirs [23]. In addition to the classification information given by FCM, the other obtained fuzzy information such as the fuzzy memberships of data can also effectively help many data mining tasks [17,24–29]. Mohammadhassani used the fuzzy information of data to build an adaptive neuro-fuzzy inference system for the prediction of shear strength of high strength concrete (HSC) beams without stirrups [24]. Nikolić utilized an adaptive neuro-fuzzy inference system (ANFIS) to estimate power coefficient and power extraction of the wind turbines and wind farm efficiency [25–27]. Nikolic proposed a neuro-fuzzy approach to detect the most important variables affecting wind speed [28,29]. The above-mentioned studies give insights on the availability and potential of FCM for engineering analysis. To achieve better clustering performance, prior information is often utilized to achieve better clustering performance. Mostly used prior information is labeled data. However, labeling data might not be available in some engineering practices since labeling data is very costly and often impossible. For instance, in TBM, the geological investigation of a tunnel is costly. In addition, the geological conditions at some specific places can be known from the geological investigation report on the ground, but they are difficult to label on the operation data generated by the TBM because of the TBM underground location problem [30]. Therefore, it is necessary to use other information instead of labeled data to improve the clustering performance of FCM. For engineering data, the spatial distributions of some attributes are similar but their functional relationship varies greatly under different operation states or working conditions. The attribute functional relationship is also a conclusion of the knowledge from previous engineering projects, which will improve the mining results if they are well handled. It is highly desirable to incorporate attribute functional relationship to FCM to improve the clustering performance of engineering data.

In this paper, we first designed the clustering objective function in which the attribute functional relationship and the data spatial distance are both considered. Then, the sequential relationship of data is introduced into this clustering objective function, and an alternating optimization method is designed to optimize the objective function to obtain the optimal segmentation results. The proposed segmentation approach is used to segment a real TBM multivariate time series data collected from a tunnel in China to show its efficiency and advantage. The technical details of the proposed approach are introduced in Section 2. Section 3 presents the application of the proposed approach to a real TBM multivariate time series data collected from a tunnel in China, and the comparison with other common time series segmentation approaches. In Section 4, TBM performance prediction is conducted to show the effectiveness and necessity of time series segmentation in engineering data analysis. The last section concludes this work.

2. Proposed time series segmentation approach

2.1 Notations and problem formulation

A time-series $T = \{x_k\}_{1 \leq k \leq n}$ is a finite set of $n$ data labeled by time points $t_1, \ldots, t_n$, where $x_k = [x_{1,k}, x_{2,k}, \ldots, x_{s,k}]^T$. A segment of $T$ is a set of consecutive time points denoted by $\text{Seg}(a,b) = \{a \leq k \leq b\}, x_a, x_{a+1}, \ldots, x_b$. The c-segmentation of time-series $T$ is a partition of $T$ to $c$ non-overlapping segments $\text{Seg}_c^T = \{\text{Seg}_i(a_i,b_i)|1 \leq i \leq c\}$, such that $a_1 = 1, b_c = n$, and $a_i = b_{i-1} + 1$. For instance, a c-segmentation partition a time series $T$ to $c$ disjoint time intervals by segment boundaries $s_1 < s_2 < \cdots < s_c$. The segmentation problem can be defined as a constrained clustering task: a data set is partitioned so that data points in a segment or a cluster are more similar in patterns than those within different clusters, but with the constraint that all data in a cluster must come from successive time series. In the following subsection, a new time series segmentation approach will be presented based on fuzzy c-means clustering.

2.1. Proposed approach

FCM is a well-known clustering approach to cluster data in unsupervised learning [19]. It partitions a given set of object data $\{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^{d \times n}$ into $c$ fuzzy clusters by minimizing an objective function $J(U, V)$ as follows:

$$J(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m D^2(x_k, v_i)$$

where $x_k = [x_{1,k}, x_{2,k}, \ldots, x_{s,k}]^T$ is an object datum, and $x_{j,k}$ is the $j$-th attribute value of $x_k; v_i$ is the $i$-th cluster prototype; $m$ is a fuzzification parameter, $m \in (1, \infty)$; $D^2(x_k, v_i)$ represents the distance metric between the $i$-th cluster prototype $v_i$ and
datum \( x_i \); \( u_{ik} \) is the membership that represents the degree to which \( x_i \) belongs to the \( i \)-th cluster and satisfies the following condition:

\[
\sum_{i=1}^{c} u_{ik} = 1 \quad (k = 1, 2, \ldots, n; \forall i, k : \; u_{ik} \in [0, 1])
\]

(2)

Define the partition matrix \( U = [u_{ik}] \in \mathbb{R}^{c \times n} \), and the matrix of cluster prototypes \( V = [v_{ij}] = [v_1, v_2, \ldots, v_t]^T \in \mathbb{R}^{c \times t} \); When the sequence relationship is considered, \( (1) \) results in the following equation:

\[
\text{Cost}\left(U, V, \text{Seg}_T^T\right) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik} \beta_i(t_k) D^2(x_i, v_i)
\]

(3)

where \( D^2(x_i, v_i) \) represents the distance between the metric prototype in the \( i \)-th segment \( v_i \) and datum \( x_i \); and \( \beta_i(t_k) \) stands for the time membership function of the \( k \)-th datum in the \( i \)-th segment. In this work, \( L_2 \)-norm is used to represent \( \beta_i(t_k) \) as follows:

\[
\beta_i(t_k) = \| t_k - v_i \|^2_2
\]

(4)

where \( v_i \) is the prototype in the \( i \)-th segment, \( t_k \) is time point of datum \( x_k \), and define the matrix of time prototypes as \( V_t \). Before segmentation, the time is \( 0 \rightarrow 1 \) normalized to \([0, 1]\). The segmentation objective function is changed as follows:

\[
\text{Cost}\left(U, V, \text{Seg}_T^T\right) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik} \| t_k - v_i \|^2_2 D^2(x_i, v_i)
\]

(5)

In engineering practice, the attribute functional relationship often exists. Therefore, incorporating such prior information is quite necessary and is expected to enhance segmentation performance. In this paper, a linear model is employed to describe the attribute functional relationship for TBM data, which is given as follows:

\[
y_{il} = b_{il}x_i + d_{il} (i = 1, 2, \ldots, c; l = 1, 2, \ldots, f)
\]

(6)

where \( y_{il} \) and \( x_i \) are the attributes of \( l \)-th functional relationship in \( i \)-th cluster of the data set; \( b_{il} \) and \( d_{il} \) are the regression coefficients of \( l \)-th relationship. Let matrices of cluster regression coefficients \( B = [b_{il}] = [b_1, b_2, \ldots, b_c] \in \mathbb{R}^{c \times c} \) and \( D = [d_{il}] = [d_1, d_2, \ldots, d_c] \in \mathbb{R}^{c \times c} \). Thus, \( D^2(x_i, v_i) \) in (3) can be reformulated as follows:

\[
D^2(x_i, v_i) = \left( \| x_i - v_i \|^2_2 + \theta \sum_{i=1}^{c} \sum_{k=1}^{n} \left( \sum_{l=1}^{f} (y_{il} - b_{il}x_i - d_{il})^2 \right) \right)
\]

(7)

where \( \theta \) is a positive weight parameter named as the gain factor. There are two parts: the first part is the distance between the cluster prototype (metric prototype) in the \( i \)-th segment \( v_i \) and datum \( x_i \) as shown in \( (1) \), which is to cluster the data according to the distance between an object datum and its cluster prototype; the second part is the residual sum of squares, which is to cluster the data according to the pairwise functional relationships among attributes. Substituting (7) into (5), the final segmentation objective function of the proposed approach can be obtained:

\[
\text{Cost} = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik} \| t_k - v_i \|^2_2 \left( \| x_i - v_i \|^2_2 + \theta \sum_{i=1}^{c} \sum_{k=1}^{n} \left( \sum_{l=1}^{f} (y_{il} - b_{il}x_i - d_{il})^2 \right) \right)
\]

(8)

**Proposition 1.** If data clustering is conducted by minimizing the objective function in (8) with the constraint (2), the necessary conditions that should be satisfied by cluster metric prototypes, cluster time prototypes, memberships, and regression coefficients lead to the following equations:

\[
v_i = \frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m}
\]

(9)

\[
v_i = \frac{\sum_{k=1}^{n} u_{ik}^m t_k}{\sum_{k=1}^{n} u_{ik}^m}
\]

(10)

\[
u_{ik} = \frac{\sum_{k=1}^{n} \left( \frac{\| t_k - v_i \|^2_2}{\| t_k - v_i \|^2_2 + \theta \sum_{k=1}^{n} (y_{ik} - b_{ik}x_i - d_{ik})^2} \right)^{\frac{-n}{2}}}{\sum_{k=1}^{n} \left( \frac{\| t_k - v_i \|^2_2}{\| t_k - v_i \|^2_2 + \theta \sum_{k=1}^{n} (y_{ik} - b_{ik}x_i - d_{ik})^2} \right)^{\frac{-n}{2}}}
\]

(11)
\[ b_{ij} = \frac{\sum_{k=1}^{m} u_{ik}^{m} (x_{ik} y_{jk} - d_{il} x_{lk})}{\sum_{k=1}^{m} u_{ik}^{m} x_{ik}^2} \]  
(12)

\[ d_{ij} = \frac{\sum_{k=1}^{m} u_{ik}^{m} (y_{jk} - b_{il} x_{lk})}{\sum_{k=1}^{m} u_{ik}^{m}} \]  
(13)

**Proof.** Applying the method of Lagrange multipliers for finding an optimum solution of the objective function in (8), we consider the following Lagrange function:

\[ L = \sum_{i=1}^{c} \sum_{k=1}^{n} \sum_{l=1}^{m} \left( \|x_k - v_l\|^2_2 + \theta \sum_{i=1}^{c} \sum_{k=1}^{n} \left( \sum_{l=1}^{m} (y_{jk} - b_{il} x_{lk} - d_{il})^2 \right) \right) + \sum_{k=1}^{n} \lambda_k \left( \sum_{i=1}^{c} u_{ik} - 1 \right) \]  
(14)

where \( \lambda_k \) is Lagrange multiplier. Fixing the parameters \( u_{ik}, b_{il}, \) and \( d_{il} \), the objective function attains its minimum when

\[ \frac{\partial L}{\partial v_l} = -2 \sum_{k=1}^{n} u_{ik}^{m} (x_k - v_l) = 0 \]  
(15)

\[ \frac{\partial L}{\partial t_l} = -2 \sum_{k=1}^{n} u_{ik}^{m} (t_k - v_l) = 0 \]  
(16)

Thus, (9) & (10) can be derived directly. For fixed parameters \( v_l, t_l, b_{il} \) and \( d_{il} \), the objective function attains its minimum when

\[ \frac{\partial L}{\partial u_{ik}} = \sum_{i=1}^{c} u_{ik} - 1 = 0 \]  
(17)

\[ \frac{\partial L}{\partial u_{ik}} = m u_{ik}^{m-1} \left( \|x_k - v_l\|^2_2 + \theta \sum_{i=1}^{c} \sum_{k=1}^{n} \left( \sum_{l=1}^{m} (y_{jk} - b_{il} x_{lk} - d_{il})^2 \right) \right) + \lambda_k = 0 \]  
(18)

Thus, we have

\[ u_{ik} = \left( - \frac{\lambda_k}{m \left( \|x_k - v_l\|^2_2 + \theta \sum_{i=1}^{c} \sum_{k=1}^{n} \left( \sum_{l=1}^{m} (y_{jk} - b_{il} x_{lk} - d_{il})^2 \right) \right) \right)^{\frac{1}{m-1}} \]  
(19)

Substituting (19) into (18),

\[ \left( - \frac{\lambda_k}{m} \right)^{\frac{1}{m-1}} \sum_{i=1}^{c} \left( \frac{1}{\left( \|x_k - v_l\|^2_2 + \theta \sum_{i=1}^{c} \sum_{k=1}^{n} \left( \sum_{l=1}^{m} (y_{jk} - b_{il} x_{lk} - d_{il})^2 \right) \right) \right)^{\frac{1}{m-1}} = 1 \]  
(20)

\[ \left( \frac{\lambda_k}{m} \right)^{\frac{1}{m-1}} = \left( \sum_{i=1}^{c} \left( \frac{1}{\left( \|x_k - v_l\|^2_2 + \theta \sum_{i=1}^{c} \sum_{k=1}^{n} \left( \sum_{l=1}^{m} (y_{jk} - b_{il} x_{lk} - d_{il})^2 \right) \right) \right)^{\frac{1}{m-1}} \right)^{-1} \]  
(21)

Substitute (21) into (19), the membership \( u_{ik} \) can be expressed as shown in (11). For fixed parameters \( u_{ik}, v_l, t_l, \) and \( d_{il} \), the objective function attains its minimum when

\[ \frac{\partial L}{\partial b_{il}} = -2 \sum_{k=1}^{n} u_{ik}^{m} x_{ik} (y_{jk} - b_{il} x_{lk} - d_{il}) = 0 \]  
(22)

Thus, (12) can be derived directly. For fixed parameters \( u_{ik}, v_l, t_l, \) and \( b_{il} \), the objective function attains its minimum when

\[ \frac{\partial L}{\partial d_{il}} = -2 \sum_{k=1}^{n} u_{ik}^{m} (y_{jk} - b_{il} x_{lk} - d_{il}) = 0 \]  
(23)

Finally, (13) can be derived directly. Based on the previously presented building blocks, the optimal segmentation results are obtained as follows:

Step 1. Choose \( m, c, \theta \) and two threshold value \( \epsilon_{\mu}, \epsilon_{\text{cost}} > 0 \). Initialize partition matrix \( U^{(0)} \) and two matrices of regression coefficients \( B^{(0)} \) and \( D^{(0)} \).
Step 2. When the iteration index is \( g (g = 1,2, \ldots) \), calculate the matrix of cluster prototypes \( V^g \) using (9), (10) and \( U^{g-1} \);
Step 3. Update two matrixes of regression coefficients \( B^g \) and \( D^g \) using (8), (9), \( U^g \) and \( V^g \);
Step 4. Update partition matrix \( U^g \) using (8), \( V^g \) and \( B^g \) and \( D^g \);
Step 5. If \( \forall i,k: \max |u^g_{ik} - u^{g-1}_{ik}| < \varepsilon \mu \) or \( |\text{Cost}^g - \text{Cost}^{g-1}| < \varepsilon \text{cost} \), then stop and obtain partition matrix \( U \), matrix of cluster prototypes \( V \), vector of time prototypes \( V_t \) and two matrixes of regression coefficients \( B \) and \( D \); otherwise set \( g = g + 1 \) and return to Step 2.

The number of segments \( c \) plays a decisive role in determining the segmentation accuracy. Therefore, it is very important to select it appropriately. In the proposed approach, the experiments with difference numbers of segments are conducted first successively, where \( c \) is suggested as 5, 10, 20, 30, 40, \ldots increasing. One popular cluster validity index, partition coefficient (PC), is used to evaluate the clustering result of each experiment as follows [31].

\[
PC = 1 - \frac{C}{c-1} \left( 1 - \frac{1}{n} \sum_{i=1}^{c} \sum_{k=1}^{n} \mu_{ik}^2 \right)
\]

(24)

The higher the PC is, the higher the cluster validity. In each experiment, its corresponding partition coefficient \( PC_{\text{current}} \) is calculated and compared with that of past experiment \( PC_{\text{past}} \). If \( PC_{\text{current}} \) is smaller than \( PC_{\text{past}} \), the real number of segments are considered to be between the current number of segments \( c_{\text{current}} \) and the past current number of segments \( c_{\text{past}} \). To ensure the reliability of the segmentation results, a relatively large current number of segments higher than \( c_{\text{current}} \) is assigned as the initial number of segments \( c_{\text{initial}} \). After that, the segmentation of \( c_{\text{initial}} \) is performed, and the spatial distance between two adjacent segments is evaluated. The two clusters are merged if they are found to be compatible. The detailed procedure of the proposed approach is formulated as follows:

Step 1. Segmenting the initial data into \( c \) clusters using the proposed approach;
Step 2. Computing the distance between each pair of adjacent cluster prototypes;
Step 3. Merging the adjacent segments according to the criterion given by the user.

The fuzzy membership \( \mu_{i,\text{merged}} \) of the resulting \( i \)-th segment is the sum of its individual segments

\[
\mu_{i,\text{merged}} = \sum_{j=1}^{\text{indi}} \mu_{ij} \quad (j = 1,2, \ldots, \text{indi})
\]

(25)

where \( \text{indi} \) is the number of individual segments. And the metric and time prototypes of \( i \)-th segment can be obtained using (9) and (10).

For instance, a time series data is first segmented as ten segments, and the distance matrix of segment metric prototypes can be obtained, as shown in Fig. 1 (left panel), where the background-filled cells represent the distance between the adjacent segments. The blue ones indicate that their values are smaller than the criterion, and the red ones are higher. It can be found the red ones are 3-4(4-3), 7-8(8-7) and 8-9(9-8). Thus, the merging results are as 1-3, 4-7, 8 and 9-10, as shown in Fig. 1 (right panel). Then, the fuzzy membership, metric prototypes and time prototypes of the merged segments can be obtained as discussed above. As this formulation of the proposed approach shows, although there is a need to define some parameters before its application, it is still possible to apply the approach for other complex engineering systems with:

![Fig. 1. Illustration of segment merging with initial distance matrix.](image-url)
multivariate time series even if almost nothing is known about the structure of these time series in advance. The results of
the segmentation should be evaluated by human experts or by the performances of other modeling and data mining tools
based on the segmented data, and if it is needed, the "knowledge worker" should return to the segmentation task with a new
set of these parameters [32–34].

3. Application to TBM time series data

3.1. Project overview

In this section, the proposed approach is applied to the segmentation of the real TBM time series, and the results are com-
pared with other time series segmentation approaches. The TBM time series used here are from a tunnel interval in China
with a diameter of 6.4 m. From the ground surface to the tunnel floor, there exist various geological layers, such as clay, rand
and rock with uneven distribution. A portion of geological characteristics of these layers are listed in Table 1. From this table,

Table 1

<table>
<thead>
<tr>
<th>Layer Type</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clay</td>
<td>...</td>
</tr>
<tr>
<td>Sand</td>
<td>...</td>
</tr>
<tr>
<td>Rock</td>
<td>...</td>
</tr>
</tbody>
</table>

it can be seen the properties of different geological layers and their space distribution varies greatly during the construction.

An earth pressure balance TBM was used to excavate the tunnel, which consists of numerous subsystems, such as cutterhead,
chamber, screw conveyor, and tail skin and auxiliaries. The operators reflect that the torque of cutterhead is correlated with
the penetration rate, but they cannot provide accurate information. It satisfies the assumption of the proposed approach, and
the proposed approach is used to segment the time series in the following work.

3.2. Data description

The TBM time series is composed of 53 attributes Table A1 (for details, see Appendix A) that were continuously measured
with a frequency of 1 Hz. The excavation process of the tunnel is divided into 1380 rings with each one representing a 1.5 m
length of the tunnel. To validate the proposed approach, five rings’ time series are randomly selected. Four popular time
series segmentation approaches, dynamic programming approach (DP) [35,36], Geva-Gath clustering-based time series seg-
mentation approach (GGC) [37], recursive binary segmentation approach (RBS) [38] and group fused LARS approach (GFLars)
[39] are compared with the proposed approach. The results are as follows.

3.3. Case 1

The duration of case 1 is 1341 s. Six main attributes representing the main states of TBM, penetration rate (V), pitch angle
(PA), rolling angle (RA), thrust of cutterhead (F), torque of cutterhead (T) and pressure of chamber at top (PCT), are shown in
Fig. 2. It can be seen that the operation mode of TBM is that the thrust of cutterhead remains constant but not the penetra-
tion rate as used in most projects [1,2]. The proposed approach is applied to the time series. To determine the initial number
of segments, the experiments with the number of segments 5, 10 and 20 are conduceted first. The experimental results are
listed in Table 2. It can be seen that the PC of five segments is higher than that of 10 segments, which indicates that the real
number of segments should be between 5 and 10.

To ensure that the relatively small segments are accurately captured, The number of initial segments c is set to 20, and the
segmentation results are shown in Fig. 3. After that, the spatial distances between each pair of two adjacent segments are
evaluated and ranked as shown in Fig. 4. From this figure, it can be found that if the selected merging criterion is relatively
high, the final number of segments would be small, which means that some important segments might not be recognized.
Otherwise, the time series data would be segmented into too many parts. To obtain a reasonable merging criterion, the
change ratio of spatial distance is proposed and defined as follows:

\[ CR = \frac{D_{i-1} - D_i}{D_i} \] (i = 2, \cdots, c_{initial} - 1) \tag{26} \]

Di is the spatial distance with rank i. Fig. 5 illustrates the change ratios of spatial distances. From this figure, it can be found
that the largest change is between D_{i=5}(0.428) and D_{i=6}(0.869). Thus, the merging criterion is suggested samller than 0.869 and
selected 0.8. The two adjacent clusters are merged if their distance is smaller than 0.8. Finally, the time series is partitioned
into nine segments as shown in Fig. 6.

The results are analyzed as follows:

Segment 1: The thrust of cutterhead increases to the set thrust. Preparing for excavation;
Segment 2: The cutterhead cuts into the material;
Segment 3: The cutterhead excavates the material steadily. It can be seen that there exists a periodic variation in the tor-
que of cutterhead and penetration rate, and the period is exactly equal to the rotation period of cutterhead. The rolling
angle and pitch angle changes linearly, which indicates there is a turn in the excavation path of TBM;
Segment 4: The penetration rate decreases first, which leads to a decrease of the cutting depth of cutters, resulting in the
decreasing of the torque of cutterhead. At the same time, the pressure of chamber increases, which means that the deslag-
Table 1
Material parameters of geological layers.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Natural density (kN/m³)</th>
<th>Natural moisture content (%)</th>
<th>Shear strength</th>
<th>Poisson ratio</th>
<th>Foundation soil coefficient (MPa/m)</th>
<th>Permeability coefficient (m/d)</th>
<th>Fak (kPa)</th>
<th>Foundation friction coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Direct shear</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cohesion (kPa)</td>
<td>Internal friction angle (°)</td>
<td>Cohesion (kPa)</td>
<td>Internal friction angle (°)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clay1</td>
<td>17.0</td>
<td>55.0</td>
<td>12</td>
<td>4.5</td>
<td>18</td>
<td>8.0</td>
<td>0.40</td>
<td>6</td>
</tr>
<tr>
<td>Clay2</td>
<td>18.5</td>
<td>–</td>
<td>0</td>
<td>18.0</td>
<td>–</td>
<td>–</td>
<td>0.35</td>
<td>8</td>
</tr>
<tr>
<td>Sand 1</td>
<td>19.0</td>
<td>30.0</td>
<td>28</td>
<td>15.0</td>
<td>30</td>
<td>16.5</td>
<td>0.32</td>
<td>18</td>
</tr>
<tr>
<td>Sand 2</td>
<td>20.5</td>
<td>–</td>
<td>0</td>
<td>32.0</td>
<td>–</td>
<td>–</td>
<td>0.22</td>
<td>18</td>
</tr>
<tr>
<td>Sand 3</td>
<td>21.0</td>
<td>–</td>
<td>0</td>
<td>35.0</td>
<td>–</td>
<td>–</td>
<td>0.25</td>
<td>25</td>
</tr>
<tr>
<td>Rock 1</td>
<td>18.5</td>
<td>32.0</td>
<td>22.5</td>
<td>20.5</td>
<td>25</td>
<td>22.5</td>
<td>0.30</td>
<td>80</td>
</tr>
<tr>
<td>Rock 2</td>
<td>19.5</td>
<td>25.5</td>
<td>35</td>
<td>27.5</td>
<td>40</td>
<td>29.5</td>
<td>0.25</td>
<td>150</td>
</tr>
<tr>
<td>Rock 3</td>
<td>19.5</td>
<td>25</td>
<td>30</td>
<td>27.0</td>
<td>35</td>
<td>30.0</td>
<td>0.25</td>
<td>90</td>
</tr>
</tbody>
</table>
ging is in poor condition. After that, the torque of cutterhead and the penetration rate increase, and the pressure of chamber decreases. Finally, the TBM enters next stable excavation process.

Segment 5: The cutterhead excavates the material steadily, similar to Segment 3
Segment 6: An adjustment process of TBM similar to Segment 4.

---

<table>
<thead>
<tr>
<th>c</th>
<th>5</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>0.8396</td>
<td>0.8391</td>
<td>0.8343</td>
</tr>
</tbody>
</table>

---

**Table 2**
Segmentation results with different numbers of segments.

---

**Fig. 2.** Time series of case 1.

**Fig. 3.** Initial segmentation of case 1.

**Fig. 4.** The spatial distance between each pair of two adjacent segments.
Segment 7: The adjustment process continues, but both pitch angle and rolling angle starts to change, and the thrust of cutterhead variates as well.
Segment 8: The cutterhead excavates the material steadily, similar to Segment 3.
Segment 9: The excavation process ends. The torque of cutterhead, thrust of cutterhead, and penetration rate decrease to zero.

The number of segments is set to nine, and the segmentation results of the other time series segmentation approaches are shown in Fig. 7. It is observed that the other time series segmentation approaches produce undesirable results. GGC does not detect the falling edge of torque of cutterhead. DP partitions the excavation preparation process into two segments but does not recognize the change of thrust around 1175 s. RBS only recognizes the first adjustment process, and GFLars cannot effectively partition the time series. The segmentation results without considering the prior information are unable to identify the changes in the operation states of TBM, which indicates the necessity and effectiveness of introducing prior information into time series segmentation in this work.

3.4. Case 2

The operation mode in this case is different from Case 1. The penetration rate of cutterhead remains relatively constant during excavation, but the thrust of cutterhead varies greatly as shown in Fig. 8. As the thrust of cutterhead is the key factor for avoiding ground settlement or uplift, it is necessary to analyze the relationship between other attributes and thrust of cutterhead under this mode to achieve reasonable thrust control, and time series segmentation can provide important information. The segmentation results of the proposed approach are shown in Fig. 9. It can be seen that the proposed approach can accurately capture three adjustment intervals. However, the approach does not identify the changes around 150 s and 350 s. Compared with the three identified adjustment intervals, the changes around 150 s and 350 s are not significant. Thus, the proposed approach still shows competitive segmentation performance. Fig. 10 shows the membership. It can be seen that the membership conversion between the last segment and the penultimate segment is relatively moderate, which indi-
Fig. 7. Comparison between the proposed approach and other approaches (case 1).

Fig. 8. Time series of case 2.
cates later time series in the penultimate segment is similar to the last segment. Similar results can be found in Fig. 9. The membership can provide additional information to the users, which is also the advantage of fuzzy c-means clustering compared with the other non-fuzzy c-means clustering-based approaches.

The segmentation number is set as the same as the proposed approach, and the results are shown in Fig. 11. It is observed GGC identifies the changes that are not recognized by the proposed approach but does not recognize the latter two adjustment intervals around 1000 s and 1300 s. The results of DP is similar to those of the proposed approach. The other two approaches (RBS and GFlars) cannot identify the main changes happen in the time series. The introduction of prior information effectively improves the segmentation results. The proposed approach provides competitive results compared with the other time series segmentation approaches.

3.5. Case 3

The time series used in this subsection is shown in Fig. 12. It can be seen that this tunnel interval is not constructed by one-time construction, but is consisted of several construction periods. The main purpose of time series segmentation in this subsection is to intelligently distinguish the initial time series into construction segments and non-construction segments. The segmentation results are shown in Fig. 13. The proposed approach is applied and the final results. It is observed that the proposed approach can effectively divide the time series into construction segments and non-construction segments. The results given by GGC are similar to the proposed approach, but the first construction interval is not segmented accurately. DP and RBS both identify wrong change points in the second non-construction interval. The results of GFlars are not acceptable. The introduction of prior information can effectively improve the segmentation results. The proposed approach obtains much better results than the other approaches tested in this work.

3.6. Case 4

This case is constructed under thrust mode (Fig. 14). There is a sudden change in the thrust of cutterhead, where it is 0 from 827 s to 893 s. Applying the proposed approach to the initial time series, the sudden change is accurately identified as shown in Fig. 15. In addition, other changes are also identified accurately. Compared with other approaches, the proposed
Fig. 11. Comparison between the proposed approach and other approaches (case 2).

Fig. 12. Time series of case 3.
approach detects all the change points, which indicates the feasibility of the proposed approach (Fig. 16). GGC accurately segments the interval with thrust zero, but the change point around 2050 s is not detected. DP, RBS, and GFlarS cannot find the interval with thrust zero. Without prior information, the segmentation approach cannot provide ideal results as well.

From the results and analysis above, it can be found that the proposed approach can provide competitive time series segmentation results. In the next section, TBM performance prediction is conducted based on the segmentation results to further show the advantages of the proposed approach.

4. TBM performance prediction based on time series segmentation

For tunnel construction using TBM, the designers and operators usually need to estimate the thrust of cutterhead to avoid the ground settlement or uplift. In this work, the thrust prediction is conducted to show the effectiveness of the proposed approach for data mining tasks. In this work, support vector regression (SVR) is used to build the thrust prediction models, and its theory and mathematical derivation can be found in Ref. [40]. In SVR, there are two main parameters, the kernel func-
tion parameter $\gamma$, and $\xi$ of the intensive-loss function. Radial Basis Function is used as the kernel function, and the parameters are set as shown in Table 3.

$$\text{Radial Basis Function} = \exp \left( -\gamma |x_i - x_j|^2 \right)$$

(27)

Two criterions, R-square ($R^2$) and Root square mean error (RMSE), are used to evaluate the prediction accuracy as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

(28)

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}}$$

(29)

where $\bar{y}_i$ is the predicted value of $y_i$, $\bar{y}$ is the mean of $y_i$, and $n$ is the number of the validation data points. The closer $R^2$ is to 1 and the smaller RMSE, the better the prediction accuracy is. Based on the segmentation results, 80% of data in each segment are randomly selected as segment training data, and the rest are used as verification data. In addition, all segment training data are combined as total training data. Based on the segment training data, the prediction models are built for each segment, respectively. Based on the total training data, only one prediction model is built.

The TBM time series data of Case 2 used, and the thrust prediction results are shown in Table 4. It can be found the results of RMSE are similar to $R^2$, thus the following discussion focuses on only one criterion $R^2$. From Table 4, it is observed that the $R^2$'s of segments 4–7 are higher than 0.900 after segmentation or not, but the prediction accuracy decreases slightly after segmentation. The data in segments 4–7 are more similar to each other compared with other segments as shown in Fig. 17. The model based on total data obtains more information from other segments, so the prediction accuracy based on total data is increased. The prediction accuracy of segments 1–2, 8 is increased greatly after segmentation, especially segment 8 (from 0.767 to 0.964). The data features of segments 1–2, 8 are different from the other segments, respectively. In the training of SVR, the model based on total data tries to reduce the prediction errors for all the training data, which results in insufficient learning for the segments with special features (such as segments 1–2, 8 in case 2). The models for each segment can
accurately capture the data feature in each segment, so the prediction accuracy improves greatly. The prediction accuracy of segment 3 decreases after segmentation, which is because there exists an adjustment process and stable excavation process (similar to segment 4) in this segment. It is difficult to accurately predict based on only the data in segment 3. The model based on total data can obtain information from other segments, so the prediction accuracy of segment 3 is higher before segmentation (Fig. 18). In addition, the overall prediction accuracy is improved after segmentation.

Table 3
Parameter settings of SVR.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>γ</td>
<td>1/data dimension</td>
</tr>
<tr>
<td>ξ</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Fig. 16. Comparison between the proposed approach and other approaches (case 4).
For some segments with relatively low prediction accuracy based on total data, the models based on segmentation results can greatly increase their prediction accuracy. The overall prediction accuracy is improved after segmentation as well. The proposed time series segmentation approach can efficiently help the data mining of TBM time series.

5. Conclusions

In this work, a new FCM-based time series segmentation approach is proposed for the TBM time series data analysis. In the proposed approach, L2-norm is used to describe the sequential or temporal relationships among the data. The common prior information of the attribute relationship is described by a linear model and introduced into the segmentation objective function to increase the segmentation accuracy. A new decision making method based on the distance between the adjacent segment prototypes is proposed to determine the required number of segments. Under this framework, the optimal segmentation results are obtained by minimizing the segmentation objective function using Langerange multiplier method. The proposed approach is applied to a real TBM time series of a tunnel in China. Compared with other time series segmentation approaches, the proposed approach is able to provide more reasonable segmentation results. The proposed approach can

<table>
<thead>
<tr>
<th>Time series segmentation</th>
<th>Seg. 1</th>
<th>Seg. 2</th>
<th>Seg. 3</th>
<th>Seg. 4</th>
<th>Seg. 5</th>
<th>Seg. 6</th>
<th>Seg. 7</th>
<th>Seg. 8</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.783</td>
<td>0.718</td>
<td>0.380</td>
<td>0.934</td>
<td>0.9637</td>
<td>0.928</td>
<td>0.964</td>
<td>0.964</td>
<td>0.797</td>
</tr>
<tr>
<td>RMSE</td>
<td>620.612</td>
<td>144.08</td>
<td>113.616</td>
<td>30.175</td>
<td>41.083</td>
<td>31.465</td>
<td>40.513</td>
<td>8.564</td>
<td>374.179</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No time series segmentation</th>
<th>Seg. 1</th>
<th>Seg. 2</th>
<th>Seg. 3</th>
<th>Seg. 4</th>
<th>Seg. 5</th>
<th>Seg. 6</th>
<th>Seg. 7</th>
<th>Seg. 8</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.677</td>
<td>0.632</td>
<td>0.461</td>
<td>0.947</td>
<td>0.996</td>
<td>0.964</td>
<td>0.985</td>
<td>0.767</td>
<td>0.761</td>
</tr>
</tbody>
</table>

Fig. 17. Time series segmentation (case 2).

Fig. 18. Prediction results of case 2.
recognize different excavation processes such as preparation process, stable excavation process and adjustment process, effectively. TBM performance prediction is conducted based on the segmentation results. The prediction accuracy of TBM performance is much improved after time series segmentation, which shows the effectiveness and necessity of time series segmentation for engineering data mining. The results can provide important references and useful information for designers and operators of TBM. This work also highlights the applicability and potential of time series segmentation in the data mining of other complex engineering systems similar to TBMs.

Acknowledgement

The research is supported by National Natural Science Foundation of China (Grant No. 51505061 and U1608256).

Appendix A

<table>
<thead>
<tr>
<th>Table A1</th>
<th>Parameters and abbreviations.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbreviation</td>
<td>Parameter</td>
</tr>
<tr>
<td>TOT</td>
<td>Temperature of oil tank (°C)</td>
</tr>
<tr>
<td>RC</td>
<td>Rotation speed of cutterhead (r/min)</td>
</tr>
<tr>
<td>FPA</td>
<td>Pressure of A group of hydraulic cylinders (bar)</td>
</tr>
<tr>
<td>FPC</td>
<td>Pressure of C group of hydraulic cylinders (bar)</td>
</tr>
<tr>
<td>PB</td>
<td>Pressure of equipment bridge (bar)</td>
</tr>
<tr>
<td>PTSTRF</td>
<td>Pressure of tail skin system at top right front (bar)</td>
</tr>
<tr>
<td>PTSTSRB</td>
<td>Pressure of tail skin system at top right back (bar)</td>
</tr>
<tr>
<td>PTSBLF</td>
<td>Pressure of tail skin system at left front (bar)</td>
</tr>
<tr>
<td>PTSTRRB</td>
<td>Pressure of tail skin system at top right back (bar)</td>
</tr>
<tr>
<td>PTSBLB</td>
<td>Pressure of tail skin system at bottom back (bar)</td>
</tr>
<tr>
<td>RSC</td>
<td>Rotation speed of screw conveyor (r/min)</td>
</tr>
<tr>
<td>TSC</td>
<td>Temperature of screw conveyor ()</td>
</tr>
<tr>
<td>RA</td>
<td>Rolling angle (°)</td>
</tr>
<tr>
<td>PCT</td>
<td>Pressure of chamber at top (bar)</td>
</tr>
<tr>
<td>PCBR</td>
<td>Pressure of chamber at bottom right (bar)</td>
</tr>
<tr>
<td>PB</td>
<td>Pressure of bentonite (bar)</td>
</tr>
<tr>
<td>GTPR</td>
<td>Grot pressure at top right (bar)</td>
</tr>
<tr>
<td>GPBL</td>
<td>Grot pressure at bottom left (bar)</td>
</tr>
<tr>
<td>PSCF</td>
<td>Pressure of screw conveyor at front (bar)</td>
</tr>
<tr>
<td>T</td>
<td>Torque of cutterhead (kNm)</td>
</tr>
<tr>
<td>RBC</td>
<td>Rotation speed of belt conveyor (m/s)</td>
</tr>
<tr>
<td>SB</td>
<td>Displacement of B group of thrust cylinders (mm)</td>
</tr>
<tr>
<td>SD</td>
<td>Displacement of C group of thrust cylinders (mm)</td>
</tr>
<tr>
<td>SABR</td>
<td>Displacement of articulated system at bottom right (mm)</td>
</tr>
<tr>
<td>SABL</td>
<td>Displacement of articulated system at bottom left (mm)</td>
</tr>
<tr>
<td>PA</td>
<td>Pitch angle (°)</td>
</tr>
</tbody>
</table>

References


