\(\theta\)-improved limited tolerance relation model of incomplete information system for evaluation of water conservancy project management modernization

Yu-qin GAO*, Guo-hua FANG, Ya-qin LIU

College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, P. R. China

Abstract: The modernization of water conservancy project management is a complicated engineering system involving a management system, a management method, management personnel, the exertion of social, economic, and ecological effects, and so on. However, indices for evaluating the modernization of water conservancy project management are usually unobtainable in practical applications. Conducting appropriate extension of the classical rough set theory and then applying it to an incomplete information system are the key to the application of the rough set theory. Based on analysis of some extended rough set models in incomplete information systems, a rough set model based on the \(\theta\)-improved limited tolerance relation is put forward. At the same time, upper approximation and lower approximation are defined under this improved relation. According to the evaluation index system and management practices, the threshold for \(\theta\) is defined. An example study indicates the practicability and maneuverability of the model.

Key words: modernization of water conservancy project management; \(\theta\)-improved limited tolerance relation model; upper and lower approximations; threshold

1 Introduction

Water conservancy project management is an endeavor of planning, coordinating, and controlling the water conservancy project in order to ensure its good operation (Gao et al. 2009). The modernization of water conservancy project management is a process of establishing a modern scientific water conservancy project management system, to meet the need of social and economic modernization and water conservancy modernization. It is a complicated engineering system involving a management system, a management method, management personnel, the exertion of social, economic, and ecologic benefits, and so on (Gu 2004). Indices for evaluating the modernization of water conservancy project management are usually unobtainable in practical applications. Therefore, evaluating the modernization of water conservancy project management objectively is a difficult problem. The rough set theory was proposed by Pawlak (1982). It has been widely used as a new mathematical tool to deal
with fuzzy and uncertain information. In the rough set theory, there is an assumption that all the obtained individual objects are completely described by the sets in the information system. In other words, the information system is complete without missing values (Pawlak 1984; Richter et al. 2003).

On account of uncertainties, it is difficult to identify an appropriate equivalence relation in the evaluation object set. Generally, when data sets are complete, the rough set theory can be used. Two main methods have been proposed to deal with the incomplete information system. The first is the pretreatment of data, such as complementing or deleting missing values. Unfortunately, such treatment more or less changes the original information system, and causes significant deviation from the results if the data completion method is not correctly used. Indirect management also increases the difficulty of data mining, and greatly reduces the efficiency. Second, the rough set theory is further extended on the basis of the compatibility relation, asymmetric relation, and tolerance relation. Yao (1996) and Kryskieewiczh (1998) used the discernibility matrix and the reduction method to study the rough set model based on the general binary relation through the definition of a field operator. Thereafter, a highly compatible piece of technology appeared: Kryskiewicz (1999) proposed a tolerance relation. Then, the non-symmetric similarity relation and valued tolerance relation are described by Stefanowski and Tsoukias (2001). Wang (2001) put forward the limited tolerance relation. Leung and Li (2003) used the concept of the maximal consistent block to describe similarity relations, and acquired knowledge from the incomplete information system. These extended rough set models weaken the equivalence relation on which the rough set theory is based, and inherit all of the basic mathematical characteristics of the classical rough set model (Pawlak 1998). Therefore, the extended rough set models have a more extensively applicable scope.

This paper presents a $\theta$-improved limited tolerance relation model through analysis of some extended rough set models in incomplete information systems, and then gives the definition of the upper and lower approximations under this improved relation. At the same time, it defines the threshold for $\theta$ according to the evaluation index system of water conservancy project management modernization and management practices. An illustrative example is analyzed to substantiate the practicability and maneuverability of the model.

2 Rough set models for incomplete information system

2.1 Rough set model based on tolerance relation

Kryskiewicz (1999) proposed the tolerance relation, and the missing values in the rough set information sheet was represented by the null values, which can be all possible values and is left out, but really exists.

An index information system is defined as $S=\langle U, A, V, f \rangle$, where $U$ is the set of evaluation objects, which is also called universe; $A = C \cup D$, with $C$ being the condition attribute set, and $D$ being the decision attribute set; $V$ is the set of the attribute values;
and \( f : U \times R \rightarrow V \) is the information function. \( B \) denotes the subset including missing index values. For \( B \subseteq A \), “*” is used to denote the value of an attribute that is missing in some cases.

Let \( U = \{x_i\} (i = 1, 2, \cdots, n) \), and \( C = \{a_k\} (k = 1, 2, \cdots, m) \), then \( \forall x_i, x_j \in U \), the tolerance relation is described as

\[
T(x_i, x_j) \iff \forall a_k \in C \left( a_k(x_i) = a_k(x_j) \lor a_k(x_i) = * \lor a_k(x_j) = * \right)
\]

(1)

Obviously, the tolerance relation is reflective and symmetric, but not necessarily transitive. \( I^T_C(x_i) \) denotes the set of evaluation objects that meet the tolerance relation in the index system \( C \), and \( I^T_C(x_i) = \{x_j | x_j \in U \land T(x_i, x_j) \} \). Based on the tolerance relation of the index system with incomplete information, \( \forall X \subseteq U \), the upper and lower approximations of \( X \) are defined respectively as

\[
\bar{X}_C^T = \{x_i | x_i \in U \land I^T_C(x_i) \cap X \neq \emptyset\}
\]

(2)

\[
\overline{X}_C^T = \{x_i | x_i \in U \land I^T_C(x_i) \subseteq X\}
\]

(3)

The tolerance relation of the index system with incomplete information regards the unknown values as potentially equal to an arbitrary known value. It readily permits two evaluation objects to be classified into the same tolerance class when there is no explicitly identical known index information, or when there is only a bit of identical known index information. Thus, the requirements of the tolerance relation are too loose to some degree.

### 2.2 Rough set model based on non-symmetric similarity relation

It is considered in the non-symmetric similarity relation model that the reason for the objects being incompletely described is not only that the knowledge is inaccurate, but also it is impossible to use all of the attributes to describe the objects. Therefore, the missing value is nonexistent rather than uncertain. Besides, it does not allow any comparison. Based on this viewpoint, the two objects are regarded as similar as long as the known attribute values of the two objects are identical.

In an incomplete information system \( S = (U, A, V, f) \), let \( U = \{x_i\} (i = 1, 2, \cdots, n) \), and \( C = \{a_k\} (k = 1, 2, \cdots, m) \), then \( \forall x_i, x_j \in U \), the non-symmetric similarity relation is described as

\[
S_C(x_i, x_j) \iff \forall a_k \in C \left( a_k(x_i) = a_k(x_j) \lor a_k(x_i) = * \lor a_k(x_j) = * \right)
\]

(4)

The non-symmetric similarity relation is reflective and transitive. \( \forall X \subseteq U \), the upper and lower approximations of \( X \) are, respectively, defined as

\[
\overline{X}_C^S = \bigcup_{i=1}^{n} S_C(x_i)
\]

(5)

\[
\underline{X}_C^S = \{x_i | x_i \in U \land S_C^{-1}(x_i) \subseteq X\}
\]

(6)

where \( S_C(x_i) = \{x_j | x_j \in U \land S_C(x_i, x_j) \} \), and \( S_C^{-1}(x_i) = \{x_j | x_j \in U \land S_C(x_i, x_j) \} \). From the model described above, if the evaluation object \( x_i \) and its non-symmetric similar object belong to \( X \), it must pertain to the class of \( X \). This is just the lower approximation. On the other
hand, if the evaluation object is non-symmetric similar to a certain object in $X$, it may pertain to the class of $X$. This is just the upper approximation. However, even though some objects that obviously have a large number of identical index data can be judged as similar intuitively, they are usually separated into different similarity classes on account of the fact that they do not satisfy the non-symmetric similarity relation. The requirements of the non-symmetric similarity relation are too strict to some degree. In a large index system, two similar evaluation objects are readily regarded as different classes for a little bit of incomplete information.

2.3 Rough set model based on limited tolerance relation

The tolerance relation is so loose that it readily causes two objects without identical known values to be classified in the same tolerance class, while the non-symmetric similarity relation is too strict and it can separate two objects that have a lot of identical known values into different similarity classes. To solve these problems, Wang (2001) proposed the limited tolerance relation.

In an incomplete information system $S = (U, A, V, f)$, let $U = \{x_i\} (i = 1, 2, \cdots, n)$, and $C = \{a_k\} (k = 1, 2, \cdots, m)$, then $\forall x_i, x_j \in U$, the limited tolerance relation is described as

$$L(x_i, x_j) \iff \forall a_k \in C (a_k(x_i) = a_k(x_j) = *) \lor \left( (P_c(x_i) \cap P_c(x_j) \neq \emptyset) \forall a_k \in \right)$$

(7)

where $P_c(x_i) = \{a_k | a_k \in C \land a_k(x_i) \neq *\}$.

The limited tolerance relation is reflective and symmetric, but not transitive. The limited tolerance class is denoted by $I^L_C(x_i) = \{x_j | x_j \in U \land L(x_i, x_j)\}$. $\forall X \subseteq U$, the upper and lower approximations are of $X$, respectively, defined as

$$X^U_C = \{x_i | x_i \in U \land I^L_C(x_i) \subseteq X\}$$

(8)

$$X^L_C = \{x_i | x_i \in U \land I^L_C(x_i) \cap D \neq \emptyset\}$$

(9)

From the model described above, the limited tolerance relation is an improvement of the tolerance relation and non-symmetric similarity relation.

3 Establishment of $\theta$-improved limited tolerance relation rough set model for incomplete information

Though the limited tolerance relation avoids classifying different evaluation objects in the same class, the objects are still considered belong to the same class when there is only one identical index value and other attributes in the information system cannot be compared. The requirement is too loose for water conservancy engineering management, which has a wide range and a large index system. To solve these problems, a new $\theta$-improved limited tolerance relation rough set model is proposed. It first sets a threshold $\theta$, which counts from zero to one. Two evaluation objects are considered belong to the same class when they meet the following...
three conditions: they satisfy the limited tolerance relation, the ratio of the number of the same
index values to the total number of the indices is not less than \( \theta \), and the ratio of the number of
different index values to the total number of the indices is not more than \( 1 - K \theta \) for \( K \geq 1 \). Second, it allows to set the threshold \( \theta \) for a better effect according to the actual requirements. Usually, there is a provision of \( 0.5 \leq \theta \leq 1 \) and \( 1 \leq K \leq 2 \) for water conservancy project management modernization evaluation. \( K \) takes a smaller value when a small part of indices are unknown, and \( K \) takes a larger value when a majority of indices are unknown.

In an incomplete information system \( S = (U, A, V, f) \), let \( U = \{x_i\} \ (i = 1, 2, \ldots, n) \), \( C = \{a_k\} \ (k = 1, 2, \ldots, m) \), and \( P_C(x_i) = \{a_k | a_k \in C \land a_k(x_i) \neq \ast\} \), then \( \forall x_i, x_j \in U \), the \( \theta \)-improved limited tolerance relation is described as

\[
\begin{align*}
\text{LR}(C, \theta) & \iff \{ (x_i, x_j) \in U \times U | |P_C(x_i) \cap P_C(x_j)|/|C| \geq \theta \} \\
& \quad \forall a_k \in C((a_k(x_i) = \ast) \lor a_k(x_j) = \ast) \rightarrow a_k(x_i) = a_k(x_j)
\end{align*}
\]

(10)

Obviously, if the ratios of the number of the known indices to the total number of the
indices for the evaluation objects \( x_i \) and \( x_j \) are both smaller than \( \theta \), \( (x_i, x_j) \in \text{LR}(C, \theta) \) indicates that all the index values of \( x_i \) and \( x_j \) are correspondingly equal. Otherwise, it indicates that the ratio of the number of the same index values of \( x_i \) and \( x_j \) to the total number of the indices is not less than \( \theta \).

\( \text{LR}(C, \theta) \) is reflective and symmetric, but not necessarily transitive. When \( \theta = 0 \), the \( \theta \)-improved limited tolerance relation is degraded into a modified tolerance relation; when \( 0 < \theta \leq 1/|C| \), we always adjust the value of \( \theta \) to make this relation more practical. We define

\[
L^\theta_C(x_i) = \{x_j \in U | (x_i, x_j) \in \text{LR}(C, \theta)\}
\]

\( L^\theta_C(x_i) \) is the maximum set of the evaluation objects that can not be easily distinguished from \( x_i \) in the index system \( C \) based on \( \theta \).

\[
\begin{align*}
U/\text{LR}(C, \theta) & = \{ L^\theta_C(x_i) | x_i \in U \}, \quad \text{and } \bigcup U/\text{LR}(C, \theta) = U .
\end{align*}
\]

In an index system with incomplete information \( S = (U, A, V, f) \), for any \( X \subseteq U \), the upper and lower approximations of \( X \) are, respectively, defined as follows:

\[
\begin{align*}
\overline{A}_\theta(X) & = \{ x_i \in U | L^\theta_C(x_i) \subseteq X \} = \bigcup \{ L^\theta_C(x_i) | x_i \subseteq X \} \\
\overline{A}_\theta(X) & = \{ x_i \in U | L^\theta_C(x_i) \subseteq X \} \subseteq X = \{ x_i \in X | L^\theta_C(x_i) \subseteq X \}
\end{align*}
\]

(11)

(12)

Similar to the complete information system, \( \overline{A}_\theta(X) \) is the set of evaluation objects that surely belong to \( X \), and \( \overline{A}_\theta(X) \) is the set of evaluation objects that possibly belong to \( X \).

4 Example study

According to China’s large and medium-sized water conservancy project management practices, an evaluation index system of water conservancy project management modernization with 34 indices was established in this study and is shown in Table 1. The implication of the indices and scoring methods can be found in Fang et al. (2009). Table 2 gathers data from ten water conservancy projects of Jiangsu Province in 2010. Data from nine water conservancy projects are complete, and those from the Taizhou Yinjiang River water
conservancy project are incomplete. \( U = \{x_1, x_2, \ldots, x_{10}\} \) is the evaluation object set, \( C = \{a_1, a_2, \ldots, a_{34}\} \) is the condition attribute set, and \( D = \{d\} \) is the decision attribute set.

**Table 1** Evaluation index system of large and medium-sized water conservancy project management modernization

<table>
<thead>
<tr>
<th>Condition attribute</th>
<th>Description</th>
<th>Condition attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_1 )</td>
<td>Rationality index of classification of water management units</td>
<td>( a_{18} )</td>
<td>Index of degree of public participation and public service of water conservancy project management</td>
</tr>
<tr>
<td>( a_2 )</td>
<td>Index of advancement and implementation degree of management and maintenance separation scheme</td>
<td>( a_{19} )</td>
<td>Perfect index of pre-arranged plan against water conservancy project accident</td>
</tr>
<tr>
<td>( a_3 )</td>
<td>Index of funded status of management personnel basic and maintenance expenditure</td>
<td>( a_{20} )</td>
<td>Perfect index of report system</td>
</tr>
<tr>
<td>( a_4 )</td>
<td>Index of advancement of management mechanism (internal performance)</td>
<td>( a_{21} )</td>
<td>Index of implementation of safety responsibility system (establishing a network with safety officers)</td>
</tr>
<tr>
<td>( a_5 )</td>
<td>Index of completion and standardized execution of water conservancy engineering checking and monitoring system</td>
<td>( a_{22} )</td>
<td>Intact rate of engineering facilities (including observation facilities)</td>
</tr>
<tr>
<td>( a_6 )</td>
<td>Index of completion and standardized execution of maintenance management system of water conservancy project</td>
<td>( a_{23} )</td>
<td>Rate of maintenance and repair</td>
</tr>
<tr>
<td>( a_7 )</td>
<td>Executive index of water conservancy project control scheme and operating system</td>
<td>( a_{24} )</td>
<td>Standardized rate of engineering design ability</td>
</tr>
<tr>
<td>( a_8 )</td>
<td>Perfect index of all kinds of plans of water conservancy project</td>
<td>( a_{25} )</td>
<td>Cleaning rate</td>
</tr>
<tr>
<td>( a_9 )</td>
<td>Perfect index of all rules and post responsibility system of public administration</td>
<td>( a_{26} )</td>
<td>Green coverage rate</td>
</tr>
<tr>
<td>( a_{10} )</td>
<td>Perfect index of talent training and scientific innovation mechanisms</td>
<td>( a_{27} )</td>
<td>Rate of treatment of water loss and soil erosion</td>
</tr>
<tr>
<td>( a_{11} )</td>
<td>Index of water conservancy project management informatization</td>
<td>( a_{28} )</td>
<td>Efficiency of engineering operation</td>
</tr>
<tr>
<td>( a_{12} )</td>
<td>Index of advancement of automatic safety monitoring system for water conservancy project</td>
<td>( a_{29} )</td>
<td>Collection ratio of water and other fees</td>
</tr>
<tr>
<td>( a_{13} )</td>
<td>Index of advancement of automatic monitoring system for brake-station project</td>
<td>( a_{30} )</td>
<td>Rate of exploitable land resources utilization</td>
</tr>
<tr>
<td>( a_{14} )</td>
<td>Index of advancement of water regime forecast and water conservancy project operating scheduling system</td>
<td>( a_{31} )</td>
<td>Rate of profit and loss of water management unit</td>
</tr>
<tr>
<td>( a_{15} )</td>
<td>Index of completion of management scope delimitation</td>
<td>( a_{32} )</td>
<td>Rate of adaptation to functional requirements of professional ability, structure and number of on-the-job personnel</td>
</tr>
<tr>
<td>( a_{16} )</td>
<td>Perfect index of management according to the laws (situation of river-related construction project approval and level of construction project management)</td>
<td>( a_{33} )</td>
<td>Proportion of personnel with education background of junior college or above</td>
</tr>
<tr>
<td>( a_{17} )</td>
<td>Index of quality construction of water politics censorial personnel</td>
<td>( a_{34} )</td>
<td>Perfect index of training plan of water conservancy project management technical personnel</td>
</tr>
</tbody>
</table>

The data in Table 2 were discretized according to the fuzzy clustering method, in which the attribution values ranging from 0 to 0.60, from 0.61 to 0.70, from 0.71 to 0.80, from 0.81 to 0.90, from 0.91 to 0.95, and from 0.96 to 1.00 were set as 0, 1, 2, 3, 4, and 5, respectively, and those larger than 1.00 were set as 6. Thus, a two-dimensional table was shaped by the attribute values. Then, an equivalent analysis was conducted on the data in the shaped table based on the \( \theta \)-improved limited tolerance relation model for incomplete information systems. Setting 0.5 for \( \theta \) and 1.0 for \( K \), the discretization universe was divided. According to the definition
of the $\theta$-improvement limited tolerance relation, if two evaluation objects satisfied the limited tolerance relation, in addition, the ratio of the number of the same index values to the total number of the indices was not less than $\theta = 0.5$, and the ratio of the number of the different index values to the total number of the indices was not more than $1 - K\theta = 0.5$, they were deemed to belong to the same class. Thus, the sample of $x_1$ and the sample of $x_2$ were equivalent according to $U/C = \{X_1, X_2, \ldots, X_9\}$, where $X_1 = \{x_1, x_2\}$, $X_2 = \{x_3\}$, $X_3 = \{x_4\}$, $X_4 = \{x_5, x_7\}$, $X_5 = \{x_6\}$, $X_6 = \{x_7, x_5, x_9\}$, $X_7 = \{x_8\}$, $X_8 = \{x_9, x_7\}$, and $X_9 = \{x_{10}\}$.

The values of condition attributes $a_5$, $a_9$, $a_{14}$, $a_{18}$, $a_{24}$, and $a_{29}$ were obtained thereby, and they are 4, 4, 4, 4, 5, and 4, respectively. Table 3 is the decision table after completion.

The information content and the degree of importance of condition attributes were calculated, and the specific calculation process is the same as that calculated with the classical rough set model, which can be found in Gao et al. (2011). Finally, the weight of each condition attribute was computed by the following formula (Slowinski 1995):

$$W_j = \frac{\text{sig}_j}{\sum_{j=1}^{34} \text{sig}_j} \quad j = 1, 2, \ldots, 34$$  \hspace{1cm} (13)

where $\text{sig}_j$ is the degree of importance of $a_j$. Table 4 shows the computational weights of
condition attributes. From Table 4, it is obvious that the weights of $a_{12}$, $a_{17}$, $a_{18}$, $a_{19}$, $a_{26}$, $a_{29}$, $a_{30}$, and $a_{33}$ are zero, which shows that these indices are redundancy indices. They have no substantial influence on evaluation results in view of practical management of the current water conservancy projects. The results are consistent with those from Gao et al. (2011), which indicates the effectiveness and maneuverability of the $\theta$-improved limited tolerance relation model.

Table 4 Computational weights of condition attributes

<table>
<thead>
<tr>
<th>Condition attribute</th>
<th>Weight</th>
<th>Condition attribute</th>
<th>Weight</th>
<th>Condition attribute</th>
<th>Weight</th>
<th>Condition attribute</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>0.0079</td>
<td>$a_{10}$</td>
<td>0.0472</td>
<td>$a_{19}$</td>
<td>0</td>
<td>$a_{28}$</td>
<td>0.0079</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0.0079</td>
<td>$a_{11}$</td>
<td>0.0551</td>
<td>$a_{20}$</td>
<td>0.0551</td>
<td>$a_{29}$</td>
<td>0</td>
</tr>
<tr>
<td>$a_3$</td>
<td>0.0079</td>
<td>$a_{12}$</td>
<td>0</td>
<td>$a_{21}$</td>
<td>0.0551</td>
<td>$a_{30}$</td>
<td>0</td>
</tr>
<tr>
<td>$a_4$</td>
<td>0.0551</td>
<td>$a_{13}$</td>
<td>0.0551</td>
<td>$a_{22}$</td>
<td>0.0079</td>
<td>$a_{31}$</td>
<td>0.0551</td>
</tr>
<tr>
<td>$a_5$</td>
<td>0.0472</td>
<td>$a_{14}$</td>
<td>0.0472</td>
<td>$a_{23}$</td>
<td>0.0551</td>
<td>$a_{32}$</td>
<td>0.0472</td>
</tr>
<tr>
<td>$a_6$</td>
<td>0.0472</td>
<td>$a_{15}$</td>
<td>0.0079</td>
<td>$a_{24}$</td>
<td>0.0472</td>
<td>$a_{33}$</td>
<td>0</td>
</tr>
<tr>
<td>$a_7$</td>
<td>0.0551</td>
<td>$a_{16}$</td>
<td>0.0551</td>
<td>$a_{25}$</td>
<td>0.0551</td>
<td>$a_{34}$</td>
<td>0.0472</td>
</tr>
<tr>
<td>$a_8$</td>
<td>0.0551</td>
<td>$a_{17}$</td>
<td>0</td>
<td>$a_{26}$</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_9$</td>
<td>0.0079</td>
<td>$a_{18}$</td>
<td>0</td>
<td>$a_{27}$</td>
<td>0.0079</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5 Conclusions

Pawlak’s rough set is not suitable for practical incomplete data because the definition of the equivalence relation in the classical rough set theory has some limitations. The extension of the classical rough set theory makes it suitable for the processing of incomplete information systems. Based on analysis of several existing extended rough set models, this study proposed a rough set model based on the $\theta$-improved limited tolerance relation for incomplete information systems. The model has four characteristics: (1) it inherits the basic mathematical characteristics from the classical rough set model, (2) it is suitable for projects in which evaluation indices are complex and cannot be obtained for some reasons, (3) it is a generalization and improvement of the rough set models based on the tolerance relation and limited tolerance relation, and (4) it allows users to adjust the threshold $\theta$ based on their experience in water conservancy project management modernization, which benefits data mining and usually leads to a better mining effect. Finally, an example study indicates the practicability and maneuverability of the model.

References


(Edited by Yan LEI)