# Pattern classification using grey tolerance rough sets

Yi-Chung Hu
*Department of Business Administration, Chung Yuan Christian University, Taoyuan, Taiwan*

## Abstract

**Purpose** – The purpose of this paper is to propose that the grey tolerance rough set (GTRS) and construct the GTRS-based classifiers.

**Design/methodology/approach** – The authors use grey relational analysis to implement a relationship-based similarity measure for tolerance rough sets.

**Findings** – The proposed classification method has been tested on several real-world data sets. Its classification performance is comparable to that of other rough-set-based methods.

**Originality/value** – The authors design a variant of a similarity measure which can be used to estimate the relationship between any two patterns, such that the closer the relationship, the greater the similarity will be.

**Keywords** Decision making, Classification, Genetic algorithm

**Paper type** Research paper

## 1. Introduction

Undoubtedly, rough set theory, which was introduced by Pawlak (1982, 1991), is very useful for analyzing vague concepts in the field of multiple attributes decision making (Bazan and Szczuka, 2001, 2005; Bazan *et al.*, 2002; Pawlak and Skowron, 2007; Pokowski, 2002; Walczak and Massart, 1999; Zhang *et al.*, 2015; Shu and Shen, 2014). Because traditional rough-set-based methods require that all quantitative attributes must be discrete (Parthaláin and Shen, 2009; Jiang and Sui, 2015), discretization is usually performed in advance. Although many discretization methods have been proposed (Bazan *et al.*, 2000; Chmielewski and Grzymala-Busse, 1996; Grzymala-Busse and Stefanowski, 2001), these methods can result in information loss. Furthermore, there is no optimal discretization method for all decision problems (Grzymala-Busse and Stefanowski, 2001). For this reason, the tolerance rough set (TRS) was further developed to handle numerical attributes effectively (Jensen and Shen, 2007; Parthaláin and Shen, 2009; Stepaniuk, 2008). In the traditional TRS, the tolerance classes are commonly determined using a simple distance measure (Skowron and Stepaniuk, 1996) along with a pre-specified similarity threshold to estimate the similarity between any two patterns distributed in feature space. The traditional TRS plays an important role in pattern classification. Indeed, it has been widely used for this purpose (e.g. Jensen and Shen, 2007; Kim and Bang, 2000; Kim, 2001; Ma and Hasi, 2005; Nguyen and Skowron, 1997; Parthaláin and Shen, 2009; Stepaniuk, 2008; Yun and Ma, 2006) by treating each class in a classification problem as a concept in a given decision table.

The concern with traditional TRS is that the simple distance measure oversimplifies the criterion aggregation because it does not consider attribute weights. Furthermore, the simple distance measure is not the only way to express the similarity between any

The author thanks the anonymous referees for their valuable comments. This research is partially supported by the Ministry of Science and Technology of Taiwan under grant MOST 104-2410-H-033-023-MY2.

Kybernetes
Vol. 45 No. 2, 2016
pp. 266-281

0368-492X
DOI 10.1108/K-04-2015-0105

two patterns. In other words, the simple distance measure for estimating proximity may not be an appropriate choice for measuring similarity for TRS. For a given pattern classification problem, because certain relationships exist between any two patterns (Hu, 2008; Hu *et al.*, 2012), it is reasonable to design a variant of a similarity measure which can be used to estimate the relationship between any two patterns, such that the closer the relationship, the greater will be the similarity.

Indeed, relationships exist between any two data sequences in the real world (Deng, 1982; Liu and Lin, 2006), although we do not know exactly what these relationships are. Unlike statistical correlation analysis, which measures the relationship between any two random variables, grey relational analysis (GRA) can find the relationships between a given reference sequence and several comparative sequences (Deng, 1982) by viewing the reference sequence as the desired goal (Hu *et al.*, 2002). This provides the motivation for implementing a relationship-based similarity measure for TRS using GRA.

This paper contributes to propose a novel grey tolerance rough set (GTRS) using GRA and applies it to pattern classification. For the GTRS, when the grey relational grade (GRG) of one pattern to another one falls above a pre-specified similarity threshold, the former can be included within the tolerance class for the latter. After grey tolerance classes for all patterns have been determined, a classification procedure can be used to assign each pattern to a class. To construct a classifier with high classification performance, because genetic algorithms are a powerful search and optimization method (Goldberg, 1989; Man *et al.*, 1999; Rooij *et al.*, 1996), a genetic-algorithm-based method has been developed here that automatically determines the relative weight of each attribute and a similarity threshold that yields high classification performance.

The rest of the paper is organized as follows. Section 2 briefly introduces rough sets and TRS with a traditional similarity measure. Using GRA, Section 3 presents the proposed GTRS, which is introduced by defining a new relationship-based similarity measure. Section 4 describes a GA-based learning algorithm for constructing the proposed GTRS-based classifier (GTRSC). Section 5 reports the experimental results of applying the proposed method to real-world data sets. Several rough-set-based classification methods are taken into account. The results show that the proposed GTRSC with subset and concept approximations performs well. Section 6 presents a discussion and conclusions.

## 2. TRSs

In this section, subsection 2.1 briefly introduces the rough set. TRS with a traditional similarity measure and variants of approximations of TRS are described in subsections 2.2 and 2.3, respectively. A classification procedure of the TRS-based classifier (TRSC) is described in subsection 2.4.

### *2.1 Rough set theory*

Rough set theory can deal with vagueness and uncertainty in decision making. Let $\mathbf{S} = (U, A \cup D)$ be a decision table, where $U$ is a non-empty set of finite elements, $A$ is a non-empty set of finite attributes, and $D$ is a non-empty set of finite decision classes. Each attribute $a \in A$ defines an information function $f_a$: $U \rightarrow V_a$, where $V_a$ is the set of values of $a$. For any $P \subseteq A$, an indiscernibility relation Ind($P$) can be defined as follows:

$$\mathrm{Ind}(P) = \left\{ (\mathbf{x}_i, \mathbf{x}_j) \in U^2 \,|\, f_i(a) = f_j(a), \forall a \in P \right\} \tag{1}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are indiscernible when $(\mathbf{x}_i, \mathbf{x}_j) \in \text{Ind}(P)$. Some equivalence classes or elementary sets are generated by Ind($P$). The elementary set of a pattern $\mathbf{x}$ is represented by $[\mathbf{x}]_P$. Any finite union of elementary sets is called a $P$-definable set (Grzymala-Busse and Siddhaye, 2004). For pattern classification, a concept $X$ consists of elements that have the same class label, so that $X \in U/D$.

Sometimes, $X \subseteq U$ is not $P$-definable. In other words, there exists elements in the same elementary set which have different class labels, so that $X$ is a vague concept. In this case, $X$ can be approximated by a pair of precise concepts (Pawlak, 1982; Walczak and Massart, 1999) using the $P$-upper approximation, $\overline{P}X$, and the $P$-lower approximation, $\underline{P}X$, as follows:

$$\overline{P}X = \left\{\mathbf{x} \middle| \mathbf{x} \in U,\ [x]_P \cap X \neq \phi\right\} \tag{2}$$

$$\underline{P}X = \left\{\mathbf{x} \middle| \mathbf{x} \in U, [x]_P \subseteq X\right\} \tag{3}$$

where $\underline{P}X \subseteq \overline{P}X$ and $\underline{P}X$ consists of elements that certainly belong to $X$, whereas $\overline{P}X$ consists of elements that possibly belong to $X$. The tuple $\langle \underline{P}X, \overline{P}X \rangle$ composed of the lower and upper approximations is called a rough set. $\underline{P}X$ and $\overline{P}X$ are so-called traditional singleton approximations. When $\overline{P}X = \underline{P}X$, $X$ is precise with respect to $P$ (i.e. $X$ is definable); when $\overline{P}X \neq \underline{P}X$, $X$ is rough with respect to $P$ (i.e. $X$ is undefinable). A vague concept has the boundary region $\text{BND}_P(X)$, consisting of elements that cannot be categorized into the concept with certainty, where $\text{BND}_P(X)$ is defined as:

$$BND_P(X) = \overline{P}X - \underline{P}X \tag{4}$$

The degree of inclusion of $\mathbf{x}$ within $X$ with respect to $P$ can be defined by a rough membership function as:

$$\mu_X^P(\mathbf{x}) = \frac{\left|[\mathbf{x}]_P \cap X\right|}{\left|[\mathbf{x}]_P\right|} \tag{5}$$

where $\mu_X^P(\mathbf{x}) \in [0, 1]$ and $|[\mathbf{x}]_P|$ denotes the cardinality of $[\mathbf{x}]_P$. Undoubtedly, the value of the rough membership function of each pattern in $\underline{P}X$ is 1, that of patterns in $\overline{P}X$ lies in the interval (0, 1), and that of patterns in $\text{BND}_P(X)$ lies in the interval (0, 1).

Because rough set theory is unable to deal with real-valued data, a discretization procedure is usually performed first. Discretization is the process of converting continuous attributes into discrete attributes. As mentioned above, discretization cannot avoid information loss. Attention has also been focussed on TRS because a TRS can handle real-valued attributes by defining a suitable similarity relation for each attribute.

### *2.2 Traditional similarity measure*

Let $\mathbf{x}_i \, R_a \, \mathbf{x}_j$ denote that $\mathbf{x}_i$ and $\mathbf{x}_j$ are similar with respect to attribute $a$, where $R_a$ is a tolerance relation with respect to attribute $a$. A standard similarity measure $S_a(\mathbf{x}_i, \mathbf{x}_j)$ with respect to $R_a$ can be defined by a simple distance function as in Skowron and Stepaniuk (1996):

$$S_a(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{\left|a(\mathbf{x}_i) - a(\mathbf{x}_j)\right|}{\max_a - \min_a} \tag{6}$$

where $a(\mathbf{x}_i)$ and $a(\mathbf{x}_j)$ are attribute values of $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively, in $V_a$, and $\max_a$ and $\min_a$ denote the maximum and minimum values respectively of the domain interval of attribute $a$. Of course, the same definition can be used for all attributes (Parthaláin and Shen, 2009). The relation between $R_a$ and $S_a(\mathbf{x}_i, \mathbf{x}_j)$ is as follows:

$$\mathbf{x}_i R_a \mathbf{x}_j \Leftrightarrow S_a(\mathbf{x}_i, \mathbf{x}_j) \geqslant \tau_a \tag{7}$$

where $\tau_a \in [0, 1]$ is the similarity threshold of attribute $a$. For $A$, an overall similarity measure $S_A(\mathbf{x}_i, \mathbf{x}_j)$ can be defined as:

$$S_A(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{a \in A} S_a(\mathbf{x}_i, \mathbf{x}_j)}{|A|} \tag{8}$$

The global tolerance relation $R_A$ is related to $S_A(\mathbf{x}_i, \mathbf{x}_j)$ as:

$$\mathbf{x}_i R_A \mathbf{x}_j \Leftrightarrow S_A(\mathbf{x}_i, \mathbf{x}_j) \geqslant \tau \tag{9}$$

where $\tau \in [0, 1]$ is a global similarity threshold based on all attributes. Unlike Ind($P$), which is an equivalence relation, a tolerance relation has the reflexive and symmetric properties but not the transitivity property.

A tolerance class $TC(\mathbf{x}_i)$ of $\mathbf{x}_i$ can be generated for a certain $\tau$ by considering the patterns that have a tolerance relation with $\mathbf{x}_i$ as:

$$TC(\mathbf{x}_i) = \{\mathbf{x}_j \in U | \mathbf{x}_i R_A \mathbf{x}_j\} \tag{10}$$

$X$ can be approximated by the lower approximation $\underline{A_\tau X}$ and the upper approximation $\overline{A_\tau X}$. As in the traditional rough set, $A_\tau X$ and $\overline{A_\tau X}$ can be defined by singletons as follows:

$$\underline{A_\tau X} = \{\mathbf{x} | \mathbf{x} \in U, TC(\mathbf{x}) \subseteq X\} \tag{11}$$

$$\overline{A_\tau X} = \{\mathbf{x} | \mathbf{x} \in U, TC(\mathbf{x}) \cap X \neq \phi\} \tag{12}$$

The tuple $\langle \underline{A_\tau X}, \overline{A_\tau X} \rangle$ is known as a TRS.

### 2.3 Variants of approximations

In addition to singletons $\underline{A_\tau X}$ and $\overline{A_\tau X}$, Hu (2015) further incorporated subset and concept approximations, introduced by Grzymala-Busse and Siddhaye (2004) for the traditional rough, set into TRS. The reason for this was that the type of approximations may have an impact on classification performance of TRSC. The subset and concept approximations for TRS can be referred to Hu (2015).

### 2.4 TRS-based classifier

In this subsection, the classification procedure of the TRSC as presented in Kim and Bang (2000) and Kim (2001) for a pattern $\mathbf{x}$ is described as follows:

- Step 1. Determine $\langle \underline{A_\tau TC}(\mathbf{x}), \overline{A_\tau TC}(\mathbf{x}) \rangle$.

  The reason for using $\underline{TC}(\mathbf{x})$ is that, $\underline{A_\tau TC}(\mathbf{x})$ consists of patterns that are certainly similar to $\mathbf{x}$, whereas $\overline{A_\tau TC}(\mathbf{x})$ consists of patterns that are possibly similar to $\mathbf{x}$. Classification information with respect to $\mathbf{x}$ can be further derived from $\underline{A_\tau TC}(\mathbf{x})$

and $\overline{A_\tau}\,TC(\mathbf{x})$. As mentioned above, for the subset and concept approximations, it is certain that $\underline{A_\tau}\,TC(\mathbf{x})$ is the same (i.e. $TC(\mathbf{x})$), but $\overline{A_\tau}\,TC(\mathbf{x})$ is not.

- Step 2. Compute the relative frequency of each decision class for $\underline{A_\tau}\,TC(\mathbf{x})$.

  If $\underline{A_\tau}\,TC(\mathbf{x})$ consists of at least two patterns, then the relative frequency of each decision class can be determined by $\underline{A_\tau}\,TC(\mathbf{x})-\{\mathbf{x}\}$. Then $\mathbf{x}$ can be assigned to the class with the largest relative frequency. The procedure can be terminated if the largest relative frequency is unique; otherwise, the class label of $\mathbf{x}$ can be determined by the boundary region $\mathrm{BND}_A(TC(\mathbf{x}))$ of $\mathbf{x}$ (i.e. $\overline{A_\tau}\,TC(\mathbf{x})-\underline{A_\tau}\,TC(\mathbf{x})$).

- Step 3. Determine the class label by the boundary region.

  Because the patterns in $\underline{A_\tau}\,TC(\mathbf{x})$ have been considered in the previous step, only the patterns in $\mathrm{BND}_A(TC(\mathbf{x}))$ contribute to the classification in this step. Let $X_i$ denote a set consisting of patterns belonging to decision class $C_i$. For a pattern $\mathbf{y}$ in $\mathrm{BND}_A(TC(\mathbf{x}))\neq\phi$, the rough membership function $\mu_{C_i}(\mathbf{y})$ for TRS with respect to $A$ can be defined as follows:

$$\mu_{C_i}(\mathbf{y}) = \frac{\left|TC(\mathbf{y}) \cap X_i\right|}{\left|TC(\mathbf{y})\right|} \tag{13}$$

  where $\mu_{C_i}(\mathbf{y})\in[0, 1]$ and $|TC(\mathbf{y})|$ denotes the cardinality of $TC(\mathbf{y})$. Then the average rough membership function of $\mathbf{x}$ with respect to $C_i$ can be computed as:

$$\overline{\mu}_{C_i}(\mathbf{x}) = \frac{1}{m}\sum_{\mathbf{y}\,\in\,\mathrm{BND}_A(\mathrm{TC}(\mathbf{x}))} \mu_{C_i}(\mathbf{y}) \tag{14}$$

  where $m$ is the number of patterns in $\mathrm{BND}_A(TC(\mathbf{x}))$. $\mathbf{x}$ can then be assigned to the class that has the largest average rough membership function. The class cannot be determined when $\mathrm{BND}_A(TC(\mathbf{x})) = \phi$.

## 3. GTRSs

### 3.1 GRA

GRA is a useful technique that can find the relationships between one major sequence and the other sequences in a given system. Let $m$ and $n$ denote the numbers of patterns and attributes, respectively. Given a reference pattern $\mathbf{x}_i = (x_{i1}, x_{i2},\ldots, x_{in})$ $(1 \leqslant i \leqslant m)$ and a set of comparative sequences. Let $\mathbf{x}_j = (x_{j1}, x_{j2},\ldots, x_{jn})$ $(1 \leqslant j \leqslant m)$ be a comparative sequence, the relationship between $\mathbf{x}_i$ and $\mathbf{x}_j$ on attribute $k$ $(1 \leqslant k \leqslant n)$ can be obtained by computing the grey relational coefficient (GRC), denoted by $\xi_k(\mathbf{x}_i, \mathbf{x}_j)$, as follows (Liu and Lin, 2006):

$$\xi_k(\mathbf{x}_i, \mathbf{x}_j) = \frac{\Delta_{\min} + \rho\Delta_{\max}}{\Delta_{jik} + \rho\Delta_{\max}} \tag{15}$$

where:

$$\Delta_{\min} = \min_s \min_l |x_{il}-x_{sl}|, 1 \leqslant s \leqslant m, 1 \leqslant l \leqslant n \tag{16}$$

$$\Delta_{\max} = \max_s \max_l |x_{il}-x_{sl}|, 1 \leqslant s \leqslant m, 1 \leqslant l \leqslant n \tag{17}$$

$$\Delta_{jik} = \left|x_{ik}-x_{jk}\right| \tag{18}$$

where $|\cdot|$ denotes the absolute value and $\rho$ is the discriminative coefficient ($0 \leqslant \rho \leqslant 1$). $\rho$ is usually specified as 0.5 (Deng, 1982), but apparently this is not an optimal setting. $\xi_k(\mathbf{x}_i, \mathbf{x}_j)$ lies in [0, 1].

The overall relationship between $\mathbf{x}_i$ and $\mathbf{x}_j$ can be obtained by further computing the grey GRG, denoted by $\Upsilon(\mathbf{x}_i, \mathbf{x}_j)$, as follows:

$$\Upsilon(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{n} w_k \xi_k(\mathbf{x}_i, \mathbf{x}_j) \tag{19}$$

where $\Upsilon(\mathbf{x}_i, \mathbf{x}_j)$ ranges from 0 to 1, and $w_k$ is the relative importance of attribute $k$.

In addition, $w_1, w_2, \ldots, w_n$ satisfy:

$$\sum_{j=1}^{n} w_j = 1 \tag{20}$$

An interesting axiom comprising four properties of the GRG can be stated as follows (Deng, 1982; Liu and Lin, 2006):

(1) Norm interval: $0 \leqslant \Upsilon(\mathbf{x}_i, \mathbf{x}_j) \leqslant 1$. This condition emphasizes that certain relationships exist between any two data sequences.

(2) Dual symmetry: $\Upsilon(\mathbf{x}_i, \mathbf{x}_j) = \Upsilon(\mathbf{x}_j, \mathbf{x}_i)$ holds when $m = 2$.

(3) Wholeness: $\Upsilon(\mathbf{x}_i, \mathbf{x}_j) \neq \Upsilon(\mathbf{x}_j, \mathbf{x}_i)$ usually holds when $m \geqslant 3$. In other words, the Dual symmetry condition is not guaranteed when $m \geqslant 3$.

(4) Approachability: $\xi_k(\mathbf{x}_i, \mathbf{x}_j)$ approaches one if $\Delta_{jik}$ approaches $\Delta_{\min}$.

### 3.2 Grey tolerance relation

Let $\mathbf{x}_i R_k^G \mathbf{x}_j$ denote that $\mathbf{x}_i$ and $\mathbf{x}_j$ are similar with respect to attribute $k$, where $R_k^G$ is a partial grey tolerance relation with respect to attribute $k$. A relationship-based similarity measure $G_k^f(\mathbf{x}_i, \mathbf{x}_j)$ with respect to $R_k^G$ can be defined as follows:

$$G_k^f(\mathbf{x}_i, \mathbf{x}_j) = w_k \xi_k(\mathbf{x}_i, \mathbf{x}_j) \tag{21}$$

The same definition can be used for all attributes. $R_k^G$ is related to $S_A^f(\mathbf{x}_i, \mathbf{x}_j)$ as follows:

$$\mathbf{x}_i R_k^G \mathbf{x}_j \Leftrightarrow G_k^f(\mathbf{x}_i, \mathbf{x}_j) \geqslant \tau_k^G \tag{22}$$

where $\tau_k^G \in [0, 1]$ is a partial similarity threshold for attribute $k$. As for $A$, an overall relationship-based similarity measure $S_A^f(\mathbf{x}_i, \mathbf{x}_j)$ can be formulated as follows:

$$S_A^G(\mathbf{x}_i, \mathbf{x}_j) = \Upsilon(\mathbf{x}_i, \mathbf{x}_j) \tag{23}$$

The global grey tolerance relation $R_A^G$ is related to $S_A^G(\mathbf{x}_i, \mathbf{x}_j)$ as follows:

$$\mathbf{x}_i R_A^G \mathbf{x}_j \Leftrightarrow S_A^G(\mathbf{x}_i, \mathbf{x}_j) \geqslant \tau_A^G \tag{24}$$

where $\tau_A^G \in [0, 1]$ is a cutoff point and a global threshold based on all attributes.

A grey tolerance class $GTC(\mathbf{x}_i)$ of $\mathbf{x}_i$ can then be generated by considering those patterns that have a grey tolerance relation with $\mathbf{x}_i$ as follows:

$$GTC(\mathbf{x}_i) = \left\{ \mathbf{x}_j \in U \middle| \mathbf{x}_i R_A^G \mathbf{x}_j \right\} \tag{25}$$

Obviously, the larger $\Upsilon(\mathbf{x}_i, \mathbf{x}_j)$ becomes, the more possible it is that $\mathbf{x}_j$ can be included within $GTC(\mathbf{x}_i)$. The lower and upper approximations of $X$, denoted by $\underline{A_{\tau^G}}X$ and $\overline{A_{\tau^G}}X$ respectively, can be determined by various approximations as described in the previous section by replacing $\underline{A_\tau}$, $\overline{A_\tau}$, and $TC(\mathbf{x})$ with $\underline{A_{\tau^G}}$, $\overline{A_{\tau^G}}$, and $GTC(\mathbf{x})$ respectively. The tuple $\langle \underline{A_{\tau^G}}X, \overline{A_{\tau^G}}X \rangle$ is called a GTRS.

For a certain $\tau_A^G$, the fact that $\mathbf{x}_i R_A^G \mathbf{x}_j$ holds does not imply that $\mathbf{x}_j R_A^G \mathbf{x}_i$ holds because of the wholeness property. Moreover, the fact that both $\mathbf{x}_i R_A^G \mathbf{x}_j$ and $\mathbf{x}_j R_A^G \mathbf{x}_z$ $(1 \leqslant z \leqslant m)$ hold do not imply that $\mathbf{x}_i R_A^G \mathbf{x}_z$ holds. In other words, unlike TRS with a traditional similarity measure, a grey tolerance relation has the reflexive property, but not the symmetric property when there are at least three patterns, and not the transitivity property. A novel GTRSC can be further constructed by combining the proposed GTRS with the TRSC classification procedure introduced in the previous section.

### 3.3 Illustrative example

To illustrate the operation of GTRS using a relationship-based similarity measure instead of the traditional similarity measure, a small decision table with two classes ($C_1$ and $C_2$) is shown in Table I, consisting of four real-valued conditional attributes and a single decision attribute. Let $\rho$ be 1, $\tau_A^G$ be 0.60, and let the attributes have equal weight ($w_k = 1/4$, $1 \leqslant k \leqslant n$). When $\mathbf{x}_1$ is taken as the reference pattern, $\Delta_{\max}$ and $\Delta_{\min}$ are equal to 28.3 and zero, respectively. $\Upsilon(\mathbf{x}_1, \mathbf{x}_1)$ is equal to one without a doubt. As for $\Upsilon(\mathbf{x}_1, \mathbf{x}_2)$, one should pay attention to $\xi_1(\mathbf{x}_1, \mathbf{x}_2)$, $\xi_2(\mathbf{x}_1, \mathbf{x}_2)$, $\xi_3(\mathbf{x}_1, \mathbf{x}_2)$, and $\xi_4(\mathbf{x}_1, \mathbf{x}_2)$. $\xi_1(\mathbf{x}_1, \mathbf{x}_2)$ can be computed as follows:

$$\xi_1(\mathbf{x}_1, \mathbf{x}_2) = \frac{0 + 0.5 \times 28.3}{1.9 + 0.5 \times 28.3} = 0.882 \tag{26}$$

$\xi_2(\mathbf{x}_1, \mathbf{x}_2)$, $\xi_3(\mathbf{x}_1, \mathbf{x}_2)$, and $\xi_4(\mathbf{x}_1, \mathbf{x}_2)$ can be further computed as 0.887, 0.934, and 1 respectively. Therefore, $\Upsilon(\mathbf{x}_1, \mathbf{x}_2)$ can be computed as follows:

$$\Upsilon(\mathbf{x}_1, \mathbf{x}_2) = (1/4)(0.882 + 0.887 + 0.934 + 1) = 0.926 \tag{27}$$

In a similar manner as shown in Table II, $\Upsilon(\mathbf{x}_1, \mathbf{x}_3)$, $\Upsilon(\mathbf{x}_1, \mathbf{x}_4)$, $\Upsilon(\mathbf{x}_1, \mathbf{x}_5)$, and $\Upsilon(\mathbf{x}_1, \mathbf{x}_6)$ can be obtained as 0.906, 0.616, 0.557, and 0.570, respectively. Hence, $GTC(\mathbf{x}_1) = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$. $GTC(\mathbf{x}_2) = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, $GTC(\mathbf{x}_3) = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$, $GTC(\mathbf{x}_4) = \{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$, $GTC(\mathbf{x}_5) = \{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$, and $GTC(\mathbf{x}_6) = \{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$ can be determined easily using $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$, $\mathbf{x}_5$, and $\mathbf{x}_6$ as reference patterns, respectively. It can be seen that $R_A^G$ conforms to the symmetric property, but does not conform to the transitivity property for $\tau_A^G = 0.60$. For instance, both $\mathbf{x}_1 R_A^G \mathbf{x}_4$ and $\mathbf{x}_4 R_A^G \mathbf{x}_6$ hold, but $\mathbf{x}_1 R_A^G \mathbf{x}_6$ does not. The singleton

**Table I.** An example decision table

| Pattern | Conditional attribute 1 | 2 | 3 | 4 | Decision attribute |
|---|---|---|---|---|---|
| $\mathbf{x}_1$ | 51.1 | 35.2 | 14.0 | 2.0 | 1 |
| $\mathbf{x}_2$ | 53.0 | 37.0 | 15.0 | 2.0 | 1 |
| $\mathbf{x}_3$ | 50.0 | 32.1 | 12.0 | 2.0 | 1 |
| $\mathbf{x}_4$ | 52.0 | 27.0 | 39.0 | 14.6 | 2 |
| $\mathbf{x}_5$ | 59.0 | 30.0 | 42.3 | 15.0 | 2 |
| $\mathbf{x}_6$ | 56.7 | 25.4 | 39.0 | 11.0 | 2 |

approximations for each grey tolerance class are summarized in Table III. For instance, because only $GTC(\mathbf{x}_2) \subseteq GTC(\mathbf{x}_2)$, singleton $\underline{A_{\tau^G}} GTC(\mathbf{x}_2)$ is equal to $\{\mathbf{x}_2\}$. Undoubtedly, singleton $\overline{A_{\tau^G}} GTC(\mathbf{x}_3) = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$ because every flow-based tolerance class has an intersection with $GTC(\mathbf{x}_3)$.

The class labels of $\mathbf{x}_1$, $\mathbf{x}_3$, $\mathbf{x}_4$, $\mathbf{x}_5$, and $\mathbf{x}_6$ can be determined directly, because the largest relative frequencies between $C_1$ and $C_2$ for $\mathbf{x}_1$, $\mathbf{x}_3$, $\mathbf{x}_4$, $\mathbf{x}_5$, and $\mathbf{x}_6$ in $\underline{A_{\tau^G}} FTC(\mathbf{x}_1) - \{\mathbf{x}_1\}$, $\underline{A_{\tau^G}} FTC(\mathbf{x}_3) - \{\mathbf{x}_3\}$, $\underline{A_{\tau^G}} FTC(\mathbf{x}_4) - \{\mathbf{x}_4\}$, $\underline{A_{\tau^G}} FTC(\mathbf{x}_5) - \{\mathbf{x}_5\}$, and $\overline{A_{\tau^G}} FTC(\mathbf{x}_6) - \{\mathbf{x}_6\}$ respectively are unique. Besides, the class labels of $\mathbf{x}_2$ can be further determined by $\overline{A_{\tau^G}} FTC(\mathbf{x}_2) - \underline{A_{\tau}^G} FTC(\mathbf{x}_2)$ because the largest relative frequencies between $C_1$ and $C_2$ for $\mathbf{x}_2$ in $\overline{A_{\tau^G}} FTC(\mathbf{x}_2) - \{\mathbf{x}_2\}$ is not unique. Because the boundary region of $\mathbf{x}_2$ is $\{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4\}$, $\mathbf{x}_2$ can be correctly assigned to $C_1$ by comparing (1/3) $(\mu_{C_1}(x_1) + \mu_{C_1}(x_3) + \mu_{C_1}(x_4)) = 0.633$ with (1/3) $(\mu_{C_2}(x_1) + \mu_{C_2}(x_3) + \mu_{C_2}(x_4)) = 0.367$.

It is interesting to note that $R_A^G$ can violate the symmetric and transitivity properties for a certain $\tau_A^G$. For instance, for $\tau_A^G = 0.62$, $GTC(\mathbf{x}_1) = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, $GTC(\mathbf{x}_3) = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$, and $GTC(\mathbf{x}_4) = \{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$. The symmetric property is violated because $\mathbf{x}_3 R_A^G \mathbf{x}_4$ holds, but $\mathbf{x}_4 R_A^G \mathbf{x}_3$ does not. Moreover, the transitivity property is violated because both $\mathbf{x}_1 R_A^G \mathbf{x}_3$ and $\mathbf{x}_3 R_A^G \mathbf{x}_4$ hold but $\mathbf{x}_1 R_A^G \mathbf{x}_4$ does not.

## 4. Genetic-algorithm-based learning algorithm

The construction of the proposed GTRSC does not involve any complex mechanisms for tuning its parameter specifications. Its construction involves basic genetic operations including selection, crossover, and mutation. To construct a GTRSC with high classification power, a real-valued GA was used to determine the relative weights of the respective attributes (i.e. $w_1$, $w_2$,…, $w_n$), the discriminative coefficient (i.e. $\rho$) and a global similarity threshold (i.e. $\tau_A^G$). In other words, $n$+2 parameters can be determined by the GA. These parameter specifications cannot be easily determined in advance by decision-makers.

| Pattern | GRC $\xi_1(\mathbf{x}_1, \mathbf{x}_j)$ | $\xi_2(\mathbf{x}_1, \mathbf{x}_j)$ | $\xi_3(\mathbf{x}_1, \mathbf{x}_j)$ | $\xi_4(\mathbf{x}_1, \mathbf{x}_j)$ | GRG |
|---|---|---|---|---|---|
| $\mathbf{x}_1$ | 1.0 | 1.0 | 1.0 | 1.0 | 1.000 |
| $\mathbf{x}_2$ | 0.882 | 0.887 | 0.934 | 1.0 | 0.926 |
| $\mathbf{x}_3$ | 0.928 | 0.820 | 0.876 | 1.0 | 0.906 |
| $\mathbf{x}_4$ | 0.940 | 0.633 | 0.361 | 0.529 | 0.616 |
| $\mathbf{x}_5$ | 0.642 | 0.731 | 0.333 | 0.521 | 0.557 |
| $\mathbf{x}_6$ | 0.716 | 0.591 | 0.361 | 0.611 | 0.570 |

**Table II.** GRG calculation using $\mathbf{x}_1$ as a reference pattern ($j = 1,\ldots, m$)

| Pattern | Singleton approximation | |
|---|---|---|
| | Lower | Upper |
| $\mathbf{x}_1$ | $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ | $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$ |
| $\mathbf{x}_2$ | $\{\mathbf{x}_2\}$ | $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ |
| $\mathbf{x}_3$ | $\{\mathbf{x}_1, \mathbf{x}_3\}$ | $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$ |
| $\mathbf{x}_4$ | $\{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$ | $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$ |
| $\mathbf{x}_5$ | $\{\mathbf{x}_5, \mathbf{x}_6\}$ | $\{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$ |
| $\mathbf{x}_6$ | $\{\mathbf{x}_5, \mathbf{x}_6\}$ | $\{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$ |

**Table III.** Singleton approximations for each grey tolerance class.

**Algorithm: construction of the proposed GTRSC**

**Input:** A set of training patterns, Population size ($n_{size}$); Total number of generations ($n_{max}$); Number of elite chromosomes ($n_{del}$, where $0 \leqslant n_{del} \leqslant n_{size}$); Crossover probability ($Pr_c$); Mutation operation with probability ($Pr_m$).

**Output:** 1. Optimal values of criterion weights, the discriminative coefficient, and a global flow-based similarity threshold;
2. A GTRSC with a high classification accuracy rate.

**Step 1. Population initialization**

Generate an initial population of $n_{size}$ chromosomes. $n_{size}$ chromosomes make up a population, and $n$+2 real-valued parameters constitute a chromosome. Randomly assign a real value ranging from zero to one to each parameter in a chromosome.

**Step 2. Chromosome evaluation**

Each chromosome corresponds to a GTRSC. The number of correctly classified training patterns is used as the fitness function. To evaluate a chromosome, the procedure for obtaining its classification accuracy rate is as follows:

(1) Compute the overall relationship-based similarity measure between any two patterns using GRA;
(2) Determine a grey tolerance class for each pattern;
(3) Generate lower and upper approximations for each grey tolerance class;
(4) Output the number of correctly classified training patterns.

**Step 3. New chromosome generation**

Let $P_k$ denote the population generated in generation $k$ ($1 \leqslant k \leqslant n_{max}$). Chromosome $i$ ($1 \leqslant k \leqslant n_{size}$) in $P_k$ is represented by $w_{i1}^k w_{i2}^k \ldots w_{in}^k \rho_i^k \tau_i^k$. Generate new chromosomes in the next generation (i.e. $P_{k+1}$) by selection, crossover, and mutation (Goldberg, 1989; Man *et al.*, 1999; Rooij *et al.*, 1996).

(1) Selection: Using a binary tournament, two chromosomes are randomly drawn from the current population, and the one with the higher fitness is placed in a mating pool.
(2) Crossover: For randomly selected chromosomes $i$, $w_{i1}^k w_{i2}^k \ldots w_{in}^k \rho_i^k \tau_i^k$, and $j$, $w_{j1}^k w_{j2}^k \ldots w_{jn}^k \rho_j^k \tau_j^k$, where $1 \leqslant i, j \leqslant n_{size}$, the probability $Pr_c$ determines whether or not the crossover operation will be performed on any two real-valued parameters in selected parents. Two new chromosomes are generated to replace their parent strings by inserting these two new chromosomes into $P_{k+1}$. The operation is performed as follows: For chromosomes $i$ and $j$, two new chromosomes $w_{i1}^{k\prime} w_{i2}^{k\prime} \ldots w_{in}^{k\prime} \rho_i^{k\prime} \tau_i^{k\prime}$ and $w_{j1}^{k\prime} w_{j2}^{k\prime} \ldots w_{jn}^{k\prime} \rho_j^{k\prime} \tau_j^{k\prime}$ can be generated, where $w_{iw}^{k\prime} = a_w w_{iw}^k + (1-a_w) w_{jw}^k$, $w_{jw}^{k\prime} = (1-a_w) w_{iw}^k + a_w w_{jw}^k$ ($1 \leqslant w \leqslant n$), $\rho_i^{k\prime} = b\rho_i^k + (1-b)\rho_j^k$, $\rho_j^{k\prime} = (1-b)\rho_i^k + b\rho_j^k$, $\tau_i^{k\prime} = c\tau_i^k + (1-c)\tau_j^k$, $\tau_j^{k\prime} = (1-c)\tau_i^k + c\tau_j^k$. $a_w$, $b$, and $c$ are all random numbers selected from the interval [0, 1].
(3) Mutation: The mutation operation with probability $Pr_m$ is performed for each real-valued parameter of the newly generated chromosomes generated by the crossover operation. If a mutation occurs, the affected gene will be changed by adding a number randomly selected using $Pr_m$ from a specified interval (e.g. [−0.01, 0.01]) to individual parameters.

**Step 4. Elitist strategy**

Randomly remove $n_{del}$ ($0 \leqslant n_{del} \leqslant n_{size}$) newly generated chromosomes from the newly generated population. Insert $n_{del}$ chromosomes with the maximum fitness from the previous generation.

**Step 5. Termination test**

Terminate the algorithm when $n_{max}$ generations have been created; otherwise, return to Step 2.

When the stopping condition has been satisfied, the algorithm is terminated, and the best chromosome with maximum fitness value among all successive generations serves as the desired solution to examine the generalization ability of the proposed GTRSC. One advantage of the proposed GTRSC is that it is simple enough to implement as a computer program without any statistical assumptions.

## 5. Computer simulations

Through computational experiments on several real-world data sets summarized in Table IV, the generalization ability of the proposed GTRSC was examined for three types of approximations. The data sets used in the computer simulations are available from the UCI machine learning repository at www.ics.uci.edu/~mlearn/MLRepository.html. The computer programs were coded in Delphi 7.0 on a personal computer with a Pentium dual CPU, Microsoft Windows XP, 2 GB RAM, and a clock rate of 3 GHz.

This section is organized as follows. Section 5.1 presents the parameter specifications used by the genetic-algorithm-based learning algorithm. Section 5.2 reports the performance of different classification methods on real-world data sets.

### *5.1 GA parameter specifications*

A number of factors can influence GA performance, including population size and the probability of applying the genetic operators. Unfortunately, there is no optimal set of parameter specifications. Based on the principles introduced by Osyczka (2002) and Ishibuchi *et al.* (2004), the parameter specifications used in each experiment were determined as follows:

(1) $n_{size} = 50$: the most common population size is between 50 and 500 individuals. Hence, 50 individuals is an acceptable minimum size.

(2) $n_{max} = 500$: the stopping condition is specified according to the available computing time.

(3) $n_{del} = 2$: to generate less perturbation in the next generation, only a small number of elite chromosomes are used.

| Data set | Number of patterns | Number of attributes | Number of classes |
|---|---|---|---|
| Australian credit approval | 690 | 14 | 2 |
| Glass | 214 | 9 | 6 |
| Hepatitis | 155 | 19 | 2 |
| Iris | 150 | 4 | 3 |
| Pima Indian diabetes | 768 | 8 | 2 |
| Sonar | 208 | 60 | 2 |
| Statlog Heart | 270 | 13 | 2 |
| Tic-Tac-Toe Endgame | 958 | 9 | 2 |
| Vote | 435 | 16 | 2 |
| Wine | 178 | 13 | 3 |

**Table IV.** Data sets used in computer simulations

(4) $Pr_c = 1.0$, $Pr_m = 0.01$: because a larger $Pr_c$ enables wide exploration of the solution space, a larger $Pr_c$ is usually specified. However, to avoid generating excessive perturbation, a smaller $Pr_m$ should be specified.

Although these parameter specifications are somewhat subjective, experimental results show that they are acceptable.

### 5.2 Classification performance evaluation

To examine the performance of the proposed method, five-fold cross-validation (5-CV) is performed ten times independently for each data set. In practice, 5-CV divides all patterns into five equally sized and disjoint subsets. For a classification method, four subsets serve as training patterns, and the single remaining subset serves as test data. This procedure is repeated until each of the five subsets has been tested. In particular, the distribution-balanced stratified CV (DBSCV) (Zeng and Martinez, 2000) can be used to estimate the generalization accuracy of a classifier. The main difference between stratified CV (SCV) and DBSCV is that, for the former, each class is uniformly distributed among the fivefolds such that the class distribution in each fold is similar to that in the original data set; the latter further considers the distribution in feature space for each class. The main reason for using DBSCV is that it has been showed that DBSCV outperforms SCV for estimating the generalization accuracy of a classifier (Zeng and Martinez, 2000).

The classification performance of the proposed GTRSC was compared with that of several representative rough-set-based classification methods, including a hierarchical version of the LM (HLM) (Wang *et al.*, 2004) using the CaseExtract algorithm, a rule-based method with shortening optimization (RSES-O) using the Rough Set Exploration System (RSES) (Bazan and Szczuka, 2001, 2005; Bazan *et al.*, 2002), a hierarchical version of RSES-O (RSES-H) (Skowron *et al.*, 2005), and a classification algorithm called RIONA which is implemented in RSES (Bazan *et al.*, 2004). The classification performance of these methods on the ten data sets was reported by Skowron *et al.* (2006), and is summarized in Table V.

It is also interesting to compare the classification performance of TRSC and its variants using various approximations, including TRSC with singleton approximations (TRSC-SI), TRSC-SU, TRSC with concept approximations (TRSC-CO), GTRSC with singleton approximations (GTRSC-SI), GTRSC with subset approximations (GTRSC-SU), and GTRSC with concept approximations (GTRSC-CO). Flow-based TRSC (FTRSC) with

**Table V.** Classification accuracy rates (%) of different classification methods

| | Classification methods | | | | | | |
|---|---|---|---|---|---|---|---|
| Data set | HLM | RSES-H | RSES-O | RIONA | TRSC-SI | TRSC-SU | TRSC-CO |
| Australian | 92.0 | 87.0 | 86.4 | 85.7 | 87.7 | 85.9 | 87.1 |
| Glass | 71.3 | 63.4 | 61.2 | 66.1 | 64.0 | 65.7 | 68.1 |
| Hepatitis | 78.7 | 81.9 | 82.6 | 82.0 | 83.3 | 83.9 | 83.5 |
| Iris | 94.1 | 95.5 | 94.9 | 94.4 | 94.3 | 95.7 | 95.2 |
| Diabetes | 72.6 | 73.8 | 73.8 | 75.4 | 67.8 | 74.1 | 73.6 |
| Sonar | 73.7 | 75.3 | 74.3 | 86.1 | 75.2 | 74.3 | 75.0 |
| Statlog Heart | 79.0 | 84.0 | 83.8 | 82.3 | 79.0 | 82.9 | 83.3 |
| TTT | 95.0 | 99.1 | 99.0 | 93.6 | 68.7 | 82.3 | 82.3 |
| Vote | 95.4 | 96.5 | 96.4 | 95.3 | 90.8 | 93.4 | 94.0 |
| Wine | 92.6 | 91.2 | 90.7 | 95.4 | 91.8 | 93.0 | 95.3 |
| Average rank | 9.15 | 7.90 | 8.80 | 8.00 | 10.90 | 9.15 | 8.45 |

singleton (FTRSC-SI), subset (FTRSC-SU), and concept approximations (FTRSC-CO) as proposed in Hu (2015) are also examined. The test results from TRSC and its variants are summarized in Tables V and VI. It can be seen that GTRSC-CO is superior to FTRSC-CO except with the Statlog Heart data. Both GTRSC-CO and GTRSC-SU are superior to TRSC-SI, TRSC-SU, and TRSC-CO.

### *5.3 Statistical analysis*

The non-parametric Friedman test (Friedman, 1940) with the post-hoc test is used to perform statistical analysis of the classification methods described above over the ten data sets. The Friedman test ranks the classification methods for each data set separately, with the best-performing method obtaining the rank of 1, the second-best 2, and so on. In case of ties, average ranks can be assigned. Let $r_j$, $k_1$, and $k_2$ denote respectively the average rank of classification method $j$, the number of classification methods, and the number of data sets used. Let classification methods 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13 correspond to TRSC-SI, TRSC-SU, TRSC-CO, FTRSC-SI, FTRSC-SU, FTRSC-CO, GTRSC-SI, GTRSC-SU, GTRSC-CO, HLM, RSES-H, RSES-O, and RIONA, respectively. In Table V, for instance, $r_1 = 10.90$ for TRSC-SI, $r_2 = 9.15$ for TRSC-SU, and $r_3 = 8.45$ for TRSC-CO.

Under the null hypothesis that the average ranks of the classification methods are equal, the $F_F$ statistic distributed as the $F$ distribution with $k_1-1$ and $(k_1-1)(k_2-1)$ degrees of freedom can be formulated using the Friedman statistic $\chi_F^2$ (Iman and Davenport, 1980):

$$F_F = \frac{(k_2-1)\chi_F^2}{k_2(k_1-1)-\chi_F^2} \tag{28}$$

where $\chi_F^2$ is defined as:

$$\chi_F^2 = \frac{12k_2}{k_1(k_1+1)}\left[\sum_{j=1}^{k_1} r_j^2 - \frac{k_1(k_1+1)^2}{4}\right] \tag{29}$$

Because $k_1 = 13$ and $k_2 = 10$, $\chi_F^2 = 51.47$. These parameters lead to $F_F = 6.759$. Because $F_F$ is greater than the critical value $F(9, 108)$ (i.e. 1.968) at the 5 percent level, the null hypothesis is rejected.

| | Classification methods | | | | | |
|---|---|---|---|---|---|---|
| Data set | FTRSC-SI | FTRSC-SU | FTRSC-CO | GTRSC-SI | GTRSC-SU | GTRSC-CO |
| Australian | 87.7 | 88.0 | 87.7 | 87.7 | 89.3 | 89.1 |
| Glass | 64.4 | 69.1 | 69.4 | 65.3 | 70.1 | 69.9 |
| Hepatitis | 84.3 | 85.6 | 84.3 | 87.5 | 86.0 | 87.0 |
| Iris | 96.2 | 95.7 | 96.2 | 97.3 | 96.1 | 96.3 |
| Diabetes | 74.7 | 75.7 | 75.9 | 73.9 | 76.5 | 76.0 |
| Sonar | 77.1 | 78.8 | 79.5 | 80.4 | 83.0 | 82.8 |
| Statlog Heart | 81.2 | 83.9 | 84.1 | 81.0 | 84.4 | 84.0 |
| TTT | 96.9 | 97.3 | 97.8 | 87.1 | 98.5 | 98.5 |
| Vote | 95.9 | 96.6 | 96.3 | 95.6 | 96.0 | 96.2 |
| Wine | 89.3 | 93.2 | 95.1 | 95.9 | 97.4 | 97.9 |
| Average rank | 7.55 | 4.85 | 4.45 | 6.15 | 2.75 | 2.80 |

**Table VI.** Classification accuracy rates (%) of TRSC variants

Subsequently, a *post hoc* test, the Nemenyi test (Nemenyi, 1963), is used to detect any significant differences among the classification methods. The classification performance of two classification methods is significantly different if the difference in their average ranks is not less than the critical difference *CD* at the $\alpha$ level:

$$CD = q_a \sqrt{\frac{k_1(k_1+1)}{6k_2}} \tag{30}$$

*CD* is equal to 5.36 because $q_{0.10} = 3.08$ (i.e. $\alpha = 0.10$) with respect to 13 classification methods. Therefore, the results can be summarized as follows:

(1) GTRSC-CO is significantly superior to TRSC-CO (8.45−2.80 = 5.65), TRSC-SU (9.15−2.80 = 6.35), TRSC-SI (10.90−2.80 = 8.10), RSES-O (8.80−2.80 = 6.00), and HLM (9.15−2.80 = 6.35).

(2) Although GTRSC-CO was not shown to be significantly superior to RSES-H and RIONA, the differences between GTRSC-CO and RSES-H and between GTRSC-CO and RIONA are slightly less than *CD* (7.90−2.80 = 5.10 and 8.00 −2.80 = 5.20), respectively. It can therefore be concluded that GTRSC-CO seems to outperform RSES-H and RIONA.

(3) As for GTRSC-SU, it is significantly superior to TRSC-CO (8.45−2.75 = 5.70), TRSC-SU (9.15−2.75 = 6.40), TRSC-SI (10.90−2.75 = 8.15), RSES-O (8.80 −2.75 = 6.05), and HLM (9.15−2.75 = 6.40).

(4) The differences between GTRSC-CO and RSES-H is slightly less than *CD* (7.90 −2.75 = 5.15). Moreover, the difference between GTRSC-CO and RIONA is just below *CD*, but close to it (8.00−2.75 = 5.25). It can therefore be concluded that GTRSC-SU seems to outperform RSES-H and RIONA.

(5) There is no significant difference among GTRSC-CO, GTRSC-SU, and GTRSC-SI. Both GTRSC-CO and GTRSC-SU performs better than FTRSC-SU and FTRSC-SI on eight out of ten data sets.

(6) There is no significant difference between GTRSC and FTRSC for the various approximations. Even so, both GTRSC-CO and GTRSC-SU performs better than FTRSC-CO, FTRSC-SU, and FTRSC-SI on seven out of ten data sets.

## 6. Discussion and conclusions

From the viewpoint of certain relationships indeed existing between any two patterns, unlike the TRS, which uses a simple distance measure to evaluate the proximity of any two patterns, this paper has incorporated relationships between any two patterns into a similarity measure for TRS. This paper contributes to present a novel GTRS in which the GRG is used to implement a relationship-based similarity measure which generates a tolerance class for each pattern. In addition, a classifier using the GTRS (i.e. GTRSC) with various approximations has been used for pattern classification. A GA-based learning algorithm is used to determine optimal parameter specifications that are not easily determined by users, including criterion weights, the discriminative coefficient, and the global similarity threshold.

Experimental results on several real-world data sets are encouraging in terms of the classification performance obtained by GTRSC. It seems that GTRSC-CO and GTRSC-

SU give satisfactory performance compared to the other representative rough-set-based classification methods considered. Especially, GTRSC-CO and GTRSC-SU have shown their superiority in classification performance on pattern classification problems in comparison with TRSC. Besides, GTRSC and FTRSC have their own characteristics, and the classification performance of GTRSC is comparable to that of FTRSC. For TRS, both FTRSC and RSRC with pairwise-comparison-based tables (Hu, 2013) have demonstrated the potential of improving classification performance by the strict preference relation, whereas GTRSC is capable of improving classification performance by using GRA to implement the relationship-based similarity measure. However, it is not possible to conclude which classification method is best because there is no such thing as the "best" classifier (Kuncheva, 2000). Experimental results also highlights that GTRSC can be applied to bankruptcy prediction, which is a financial decision-making problem with two classes (non-bankrupt and bankrupt firms).

In this paper, GTRS is used to construct a classifier using supervised learning. Nevertheless, it would be interesting to extend GTRSC to unsupervised clustering problems. It is obvious that the GRG is defined by an additive set function $\mu$ on all singletons $\{x_j\}$ with $\mu(\{x_j\}) = w_j$. In other words, like the weighted average method or the Lebesgue integral, the traditional GRG is an additive integral in which noninteraction among the attributes involved is assumed. Nevertheless, an assumption of additivity may not be realistic in many applications (Wang *et al.*, 1998) because the variables are not always independent. It would be very interesting to implement a nonadditive similarity measure using the nonadditive version of GRG proposed in Hu (2008). Whether nonadditive GRG can have any impact on classification performance when incorporated into GTRSC remains to be studied in future work.

## References

Bazan, J., Szczuka, M., Wojna, A. and Wojnarski, M. (2004), "On the evolution of rough set exploration system", in Tsumoto, S., Słowiński, R., Komorowski, J. and Grzymala-Busse, J.W. (Eds), *Lecture Notes in Artificial Intelligence*, Vol. 3066, Springer-Verlag, Heidelberg, pp. 592-601.

Bazan, J.G. and Szczuka, M. (2001), "RSES and RSESlib – A collection of tools for rough set computation", in Figueira, W. and Yao, Y. (Eds), *Lecture Notes in Computer Science*, Vol. 2005, Springer-Verlag, Berlin, pp. 106-113.

Bazan, J.G. and Szczuka, M. (2005), "The rough set exploration system", in James, F.P. and Andrzej, S. (Eds), *Lecture Notes in Computer Science*, Vol. 3400, Springer-Verlag, Berlin, pp. 37-56.

Bazan, J.G., Nguyen, H.S., Nguyen, S.H., Synak, P. and Wróblewski, J. (2000), "Rough set algorithms in classification problem", in Polkowski, L., Tsumoto, S. and Lin, T.Y. (Eds), *Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems*, Physica-Verlag, Heidelberg, pp. 49-88.

Bazan, J.G., Szczuka, M. and Wroblewski, J. (2002), "A new version of rough set exploration system", in Alpigini, J.J., Peters, J.F., Skowron, A. and Zhong, N. (Eds), *Lecture Notes in Computer Science*, Vol. 2475, Springer-Verlag, Heidelberg, pp. 397-404.

Chmielewski, M.R. and Grzymala-Busse, J.W. (1996), "Global discretization of continuous attributes as preprocessing for machine learning", *International Journal of Approximate Reasoning*, Vol. 15 No. 4, pp. 319-331.

Deng, J.L. (1982), "Control problems of grey systems", *Systems and Control Letters*, Vol. 1 No. 5, pp. 288-294.

Friedman, M. (1940), "A comparison of alternative tests of significance for the problem of m rankings", *Annals of Mathematical Statistics*, Vol. 11, pp. 86-92.

Goldberg, D.E. (1989), *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Boston, MA.

Grzymala-Busse, J.W. and Siddhaye, S. (2004), "Rough set approaches to rule induction from incomplete data", *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, *Perugia*, pp. 923-930.

Grzymala-Busse, J.W. and Stefanowski, J. (2001), "Three discretization methods for rule induction", *International Journal of Intelligent Systems*, Vol. 16 No. 1, pp. 29-38.

Hu, Y.C. (2008), "Nonadditive grey single-layer perceptron with Choquet integral for pattern classification problems using genetic algorithms", *Neurocomputing*, Vol. 72 Nos 1-3, pp. 332-341.

Hu, Y.C. (2013), "Rough sets for pattern classification using pairwise-comparison-based tables", *Applied Mathematical Modelling*, Vol. 37 Nos 12-13, pp. 7330-7337.

Hu, Y.C. (2015), "Flow-based tolerance rough sets for pattern classification", *Applied Soft Computing*, Vol. 27, pp. 322-331.

Hu, Y.C., Chen, R.S., Hsu, Y.T. and Tzeng, G.H. (2002), "Grey self-organizing feature maps", *Neurocomputing*, Vol. 48 No. 1, pp. 863-877.

Hu, Y.C., Wang, J.H. and Chang, C.Y. (2012), "Flow-based grey single-layer perceptron with fuzzy integral", *Neurocomputing*, Vol. 91, pp. 86-89.

Iman, R.L. and Davenport, J.M. (1980), "Approximations of the critical region of the Friedman statistic", *Communications in Statistics*, Vol. 9 No. 6, pp. 571-595.

Ishibuchi, H., Nakashima, T. and Nii, M. (2004), *Classification and Modeling with Linguistic Information Granules: Advanced Approaches to Linguistic Data Mining*, Springer, Heidelberg.

Jensen, R. and Shen, Q. (2007), "Tolerance-based and fuzzy-rough feature selection", *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ- IEEE'07)*, *London*, pp. 877-882.

Jiang, F. and Sui, Y. (2015), "A novel approach for discretization of continuous attributes in rough set theory", *Knowledge-Based Systems*, Vol. 73, pp. 324-334.

Kim, D. (2001), "Data classification based on tolerant rough set", *Pattern Recognition*, Vol. 34 No. 8, pp. 1613-1624.

Kim, D. and Bang, S.Y. (2000), "A handwritten numeral character classification using tolerant rough set", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22 No. 9, pp. 923-937.

Kuncheva, L.I. (2000), *Fuzzy Classifier Design*, Physica-Verlag, Heidelberg.

Liu, S. and Lin, Y. (2006), *Grey Information: Theory and Practical Applications*, Springer-Lerlag, London.

Ma, J. and Hasi, B. (2005), "Remote sensing data classification using tolerant rough set and neural networks", *Science in China Set. D Earth Sciences*, Vol. 48 No. 12, pp. 2251-2259.

Man, K.F., Tang, K.S. and Kwong, S. (1999), *Genetic Algorithms: Concepts and Designs*, Springer, London.

Nemenyi, P.B. (1963), "Distribution-Free Multiple Comparisons", PhD thesis, Princeton University, Princeton.

Nguyen, S.H. and Skowron, A. (1997), "Searching for relational patterns in data", in Komorowski, J. and Zytkow, J. (Eds), *Lecture Notes in Computer Science*, Vol. 1263, Springer-Verlag, Berlin, pp. 265-276.

Osyczka, A. (2002), *Evolutionary Algorithms for Single and Multicriteria Design Optimization*, Physica-Verlag, New York, NY.

Parthaláin, N.M. and Shen, Q. (2009), "Exploring the boundary region of tolerance rough sets for feature selection", *Pattern Recognition*, Vol. 42 No. 5, pp. 655-667.

Pawlak, Z. (1982), "Rough sets", *International Journal of Computer and Information Sciences*, Vol. 11 No. 5, pp. 341-356.

Pawlak, Z. (1991), *Rough Sets: Theoretical Aspects of Reasoning About Data*, Kluwer, Dordrecht.

Pawlak, Z. and Skowron, A. (2007), "Rough sets and boolean reasoning", *Information Sciences*, Vol. 177 No. 1, pp. 41-73.

Pokowski, L. (2002), *Rough Sets: Mathematical Foundations*, Physica-Verlag, Heidelberg.

Rooij, A.J.F., Jain, L.C. and Johnson, R.P. (1996), *Neural Network Training Using Genetic Algorithms*, World Scientific, Singapore.

Shu, W. and Shen, H. (2014), "Incremental feature selection based on rough set in dynamic incomplete data", *Pattern Recognition*, Vol. 47 No. 12, pp. 3890-3906.

Skowron, A. and Stepaniuk, J. (1996), "Tolerance approximation spaces", *Fundamenta Informaticae*, Vol. 27 Nos 2-3, pp. 245-253.

Skowron, A., Wang, H., Wojna, A. and Bazan, J.G. (2005), "A hierarchical approach to multimodal classification", in Slezak, D., Yao, J.T., Peters, J.F., Ziarko, W. and Hu, X. (Eds), *Lecture Notes in Artificial Intelligence*, Vol. 3642, Springer-Verlag, Berlin, pp. 119-127.

Skowron, A., Wang, H., Wojna, A. and Bazan, J.G. (2006), "Multimodal classification: Case studies", in Peters, J.F. and Skowron, A. (Eds), *Lecture Notes in Computer Science*, Vol. 4100, Springer-Verlag, Berlin, pp. 224-239.

Stepaniuk, J. (2008), *Rough Granular Computing in Knowledge Discovery and Data Mining*, Springer, Berlin.

Walczak, B. and Massart, D.L. (1999), "Rough set theory", *Chemometrics and Intelligent Laboratory Systems*, Vol. 47 No. 1, pp. 1-16.

Wang, H., Düntsch, I., Gediga, G. and Skowron, A. (2004), "Hyperrelations in version space", *International Journal of Approximate Reasoning*, Vol. 36 No. 3, pp. 223-241.

Wang, W., Wang, Z. and Klir, G.J. (1998), "Genetic algorithms for determining fuzzy measures from data", *Journal of Intelligent and Fuzzy Systems*, Vol. 6 No. 2, pp. 171-183.

Yun, O. and Ma, J. (2006), "Land cover classification based on tolerant rough set", *International Journal of Remote Sensing*, Vol. 27 No. 14, pp. 3041-3047.

Zeng, X. and Martinez, T.R. (2000), "Distribution-balanced stratified cross-validation for accuracy estimation", *Journal of Experimental and Theoretical Artificial Intelligence*, Vol. 12 No. 1, pp. 1-12.

Zhang, X., Dai, J. and Yu, Y. (2015), "On the union and intersection operations of rough sets based on various approximation spaces", *Information Sciences*, Vol. 292, pp. 214-229.

**Corresponding author**
Dr Yi-Chung Hu can be contacted at: ychu@cycu.edu.tw