International Journal of Approximate Reasoning 54 (2013) 1107–1129

Contents lists available at SciVerse ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Optimal scale selection for multi-scale decision tables

Wei-Zhi Wu [a,*], Yee Leung [b,c]

[a] School of Mathematics, Physics and Information Science, Zhejiang Ocean University, Zhoushan, Zhejiang 316004, PR China
[b] Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong
[c] Institute of Environment, Energy and Sustainability, The Chinese University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 14 July 2012
Received in revised form 28 March 2013
Accepted 28 March 2013
Available online 6 April 2013

Keywords:
Belief functions
Granular computing
Information tables
Multi-scale decision tables
Probabilistic rough set models
Rough sets

ABSTRACT

Human beings often observe objects or deal with data hierarchically structured at different levels of granulations. In this paper, we study optimal scale selection in multi-scale decision tables from the perspective of granular computation. A multi-scale information table is an attribute-value system in which each object under each attribute is represented by different scales at different levels of granulations having a granular information transformation from a finer to a coarser labelled value. The concept of multi-scale information tables in the context of rough sets is introduced. Lower and upper approximations with reference to different levels of granulations in multi-scale information tables are defined and their properties are examined. Optimal scale selection with various requirements in multi-scale decision tables with the standard rough set model and a dual probabilistic rough set model are discussed respectively. Relationships among different notions of optimal scales in multi-scale decision tables are further analyzed.

1. Introduction

Granular computing (GrC) is an approach for knowledge representation and data mining. The purpose of GrC is to seek for an approximation scheme which can effectively solve a complex problem at a certain level of granulation. The root of GrC comes from the concept of information granulation which was first introduced by Zadeh in the context of fuzzy sets in 1979 [57,58]. Since its conception, "Granular computing" has become a fast growing field of research in recent years (see e.g., [1,2,10,17,18,22,28,29,39,40,43,46,47,49,53,54,62]).

A primitive notion in GrC is called a *granule* which is a clump of objects drawn together by the criteria of indistinguishability, similarity or functionality [58]. A granule may be interpreted as one of the numerous small particles forming a larger unit. Alternatively, a granule may be considered as a localized view or a specific aspect of a large unit satisfying a given specification. The set of granules provides a representation of the unit with respect to a particular level of granularity. The construction, representation, and interpretation of granules, as well as the search for relations among granules represented as IF–THEN rules having granular variables and granular values are some of the fundamental issues of GrC. The process of constructing information granules is called *information granulation*. Granulation of a universe of discourse involves the decomposition of the universe into parts, or the grouping of individual elements or objects into classes, based on available information and knowledge [49,50].

An important and commonly used model for GrC is the partition model proposed by Yao [53]. This model is constructed by granulating a finite universe of discourse through a family of pairwise disjoint subsets under an equivalence relation. An equivalence relation allows us to model the passage from one level of detail to another, but does not, on its own, model more than two levels of details needed in practice. For example, we have maps/geographical information systems represented in multiple scales, and remotely sensed data obtained at multiple resolutions. Based on this observation and by employing the notion of labelled partition, Bittner and Smith [3] developed an ontologically-motivated formal theory of granular

* Corresponding author. Tel.: +86 580 8180928; fax: +86 580 2550029.
E-mail addresses: wuwz8681@sina.com, wuwz@zjou.edu.cn (W.-Z. Wu), yeeleung@cuhk.edu.hk (Y. Leung).

0888-613X/$ - see front matter
http://dx.doi.org/10.1016/j.ijar.2013.03.017

partitions which is relatively comprehensive and useful for granular levels, but it does not address the types of aggregation commonly used in data mining and conceptual data modelling, and it has no functions, no mechanisms to deal with multiple granulation hierarchies for different perspectives. In order to represent hierarchical structure of data measured at different levels of granularities, Keet [11] explored a formal theory of granularity to build structure of the contents for different types of granularities. More recently, Wu and Leung [42] developed a new knowledge representation system, called *multi-scale granular labelled partition structure*, in which data are represented by different scales at different levels of granulations having a granular information transformation from a finer to a coarser labelled partition.

A natural consequence of granulation is the problem of approximating concepts using granules. The theory of rough sets, proposed by Pawlak [24] has been shown to perform well in constructing a granulated view of the universe of discourse and for interpreting, representing, and processing concepts in the granulated universe. It enables us to precisely define and analyze many notions of GrC. For example, the equivalence relation in a Pawlak approximation space groups together entities which are in some sense indiscernible or similar, called equivalence classes. These equivalence classes are the basic building blocks for the representation and approximation of any subset of the universe of discourse. Based on the approximation space, the notions of lower and upper approximations of decision classes can be calculated. Using the lower and upper approximations, knowledge hidden in data set may be unravelled and expressed in the form of IF–THEN granular rules. So, rough set theory is one of the most advanced areas popularizing GrC.

The basic idea of rough set theory is the acquisition of knowledge in the form of a set of decision rules unravelled from an information table via an objective knowledge induction process. Various approaches using rough set theory have been proposed to induce decision rules from data sets taking the form of decision tables [4,6,8,13,14,19,25,30,60]. So far, in the literature, each object under each attribute in almost all information tables can only take on one value, that is, almost all information tables in the rough-set data analysis are single scale information tables. However, objects are usually measured at different scales under the same attribute [15]. In many real-life multi-scale information tables, an object can take on as many values as there are scales under the same attribute. A simple example is that the examination results of mathematics for students can be recorded as natural numbers between 0 to 100, and it can also be graded as “Excellent", “Good", “Moderate", “Bad", and “Unacceptable". Sometimes, if needed, it might be graded into two values, “Pass" and “Fail". Hence, how to discover knowledge in hierarchically organized information tables is of particular importance in real-life data mining. In [31,32], Qian et al. extended Pawlak's rough set model to the so-called multi-granulation rough set models for knowledge acquisition in the context of complete and incomplete information tables. It can be seen that the multi-granulation rough set models proposed in [31,32] are in fact obtained by adding/deleting attributes in the information tables. In [42], Wu and Leung introduced the notion of *multi-scale information tables* from the perspective of granular computation, represented the structure of and relationships among information granules, and analyzed knowledge acquisition in multi-scale decision tables under different levels of granularities. In a multi-scale information table, each object under each attribute is represented by different scales at different levels of granulations having, a granular information transformation from a finer to a coarser labelled value.

Effective information and knowledge management must facilitate zooming in or zooming out of a section of interest for diverse types of users, abstracting away details when it is not needed, and focussing on a level of detail relevant to the domain experts' information needs. In short, accessing and using information and knowledge at the optimum level of granularity must be considered [11]. So, for a given multi-scale information table, we believe that there are also two key issues crucial to the discovery of knowledge in the sense of granular IF–THEN rules. One is the optimal scale selection for choosing a proper decision table with some requirements for final decision or classification, and the other is knowledge reduction by reducing attributes in the selected decision table to maintain structure consistency for the induction of concise decision rules. Since many rough set approaches have been proposed to knowledge reduction in information tables, in the present paper we mainly focus on the first issue. To this end, we will use two rough set models, one is the standard Pawlak rough set model and the other is a dual probabilistic rough set model.

In the next section, we introduce some basic notions related to Pawlak's rough sets, information tables with granular knowledge representation, and the Dempster–Shafer theory of evidence. The concept of multi-scale decision tables with representation of granules and rough set approximations are reviewed in Section 3. In Section 4, we investigate optimal scale selection in multi-scale decision tables by employing the standard Pawlak rough set model and a dual probabilistic rough set model respectively. We then conclude the paper with a summary and outlook for further research in Section 5.

2. Basic notions related to information tables, decision tables and rough set approximations

In this section we review some basic notions of information tables and rough set approximations.

Throughout this paper, for a nonempty set U, the class of all subsets of U is denoted by $\mathcal{P}(U)$. For $X \in \mathcal{P}(U)$, we denote the complement of X in U as $\sim X$, i.e. $\sim X = U - X = \{x \in U | x \notin X\}$.

2.1. Pawlak rough set approximations

Definition 1 [24]. Let U be a finite and nonempty set called the universe of discourse. If $R \subseteq U \times U$ is an equivalence relation on U, that is, R is a reflexive, symmetric and transitive binary relation on U, then the pair (U, R) is called a Pawlak approximation space.

The equivalence relation R in a Pawlak approximation space (U, R) partitions the universe of discourse U into disjoint subsets. Such a partition of the universe of discourse is a quotient set of U and is denoted by $U/R = \{[x]_R | x \in U\}$, where $[x]_R = \{y \in U | (x, y) \in R\}$ is the R-equivalence class containing x. Elements of U/R are called elementary sets. The empty set $\emptyset$ and the union of one or more elementary sets are called definable. The equivalence relation and the induced equivalence classes or the approximation space (U, R) may be regarded as the available information or knowledge about the objects under consideration. For two elements $x, y \in U$, if $(x, y) \in R$, we say that x and y are indistinguishable. In view of granular computing, equivalence classes are the basic building blocks for the representation and approximation of any subset of the universe of discourse. Each equivalence class may be viewed as a granule consisting of indistinguishable elements, and it is also referred to as an equivalence granule.

Definition 2 [24]**.** Let (U, R) be a Pawlak approximation space. For an arbitrary set $X \in \mathcal{P}(U)$, one can characterize X by a pair of *lower and upper approximations* which are defined as follows:

$$\underline{R}(X) = \cup\{[x]_R | [x]_R \subseteq X\}, \ \overline{R}(X) = \cup\{[x]_R | [x]_R \cap X \neq \emptyset\}. \tag{1}$$

The pair $(\underline{R}(X), \overline{R}(X))$ is called the *Pawlak rough set* of X with respect to (w.r.t.) (U, R).

Evidently, the lower and upper approximations can be equivalently defined by:

$$\underline{R}(X) = \{x \in U | [x]_R \subseteq X\}, \ \overline{R}(X) = \{x \in U | [x]_R \cap X \neq \emptyset\}. \tag{2}$$

We can see from Definition 2 that X is definable if and only if $\underline{R}(X) = \overline{R}(X)$.

Given a subset $X \subseteq U$, by using the lower and upper approximations, the universe of discourse can be divided into three pair-wise disjoint regions, namely, the positive, the negative, and the boundary regions:

$\mathrm{POS}_R(X) = \underline{R}(X)$,
$\mathrm{BN}_R(X) = \overline{R}(X) - \underline{R}(X)$,
$\mathrm{NEG}_R(X) = \sim \overline{R}(X) = U - \overline{R}(X)$.

An element of the positive region $\mathrm{POS}_R(X)$ definitely belongs to X, an element of the negative region $\mathrm{NEG}_R(X)$ definitely does not belong to X, and an element of the boundary region $\mathrm{BN}_R(X)$ only possibly belongs to X.

Observing that rules constructed from the three regions are associated with different actions and decisions, by employing probabilistic rough sets and Bayesian decision theory Yao [51] proposed a new notion of three-way decision rules in which a positive rule makes a decision of acceptance, a negative rule makes a decision of rejection, and a boundary rule makes a decision of abstaining.

The lower and upper approximations can also be represented as the concept of rough membership functions defined by a conditional probability [27]. Let $P : \mathcal{P}(U) \to [0, 1]$ be a probability function defined on the power set $\mathcal{P}(U)$, and R an equivalence relation on U. The triplet (U, R, P) is called a *probabilistic approximation space* [27]. For $X \in \mathcal{P}(U)$, its *rough membership function* is given by the conditional probability as follows:

$$\mu_X(x) = P(X | [x]_R), \quad x \in U. \tag{3}$$

Rough membership value of an element belonging to X is the probability of the element in X given that the element is in $[x]_R$. For a finite universe, the rough membership function is defined by Pawlak and Skowron as follows [26]:

$$\mu_X(x) = \frac{|X \cap [x]_R|}{|[x]_R|}, \quad x \in U. \tag{4}$$

where $|X|$ denotes the cardinality of the set X.

It is easy to see that the lower and upper approximations of a set X w.r.t. (U, R) are the core and support of the fuzzy set μ_X, respectively, i.e.,

$$\underline{R}(X) = \{x \in U | \mu_X(x) = 1\}, \ \overline{R}(X) = \{x \in U | \mu_X(x) > 0\}. \tag{5}$$

The *accuracy* of rough set approximation is defined as follows [24]:

$$\alpha_R(X) = \frac{|\underline{R}(X)|}{|\overline{R}(X)|}, \tag{6}$$

where for the empty set $\emptyset$, we define $\alpha_R(\emptyset) = 1$. Clearly, $0 \leq \alpha_R(X) \leq 1$. If X is definable, then $\alpha_R(X) = 1$.

2.2. Belief structures and belief functions

The Dempster–Shafer theory of evidence, also called the "evidence theory" or the "belief function theory", is treated as a promising method of dealing with uncertainty in intelligence systems. The basic representational structure in the Dempster–Shafer theory of evidence is a belief structure [33].

Definition 3. Let U be a non-empty finite set, a set function $m : \mathcal{P}(U) \to [0, 1]$ is referred to as a basic probability assignment if it satisfies axioms (M1) and (M2):

$$\text{(M1)}\ m(\emptyset) = 0, \quad \text{(M2)}\ \sum_{A \subseteq U} m(A) = 1.$$

The value $m(A)$ represents the degree of belief that a specific element of U belongs to set A, but not to any particular subset of A. A set $A \in \mathcal{P}(U)$ with nonzero basic probability assignment is referred to as a *focal element*. We denote by $\mathcal{M}$ the family of all focal elements of m. The pair $(\mathcal{M}, m)$ is called a *belief structure* on U.

Associated with each belief structure, a pair of belief and plausibility functions can be defined [33].

Definition 4. Let $(\mathcal{M}, m)$ be a belief structure on U. A set function $\mathrm{Bel} : \mathcal{P}(U) \to [0, 1]$ is referred to as a belief function on U if

$$\mathrm{Bel}(X) = \sum_{A \subseteq X} m(A), \ \forall X \in \mathcal{P}(U). \tag{7}$$

A set function $\mathrm{Pl} : \mathcal{P}(U) \to [0, 1]$ is referred to as a plausibility function on U if

$$\mathrm{Pl}(X) = \sum_{A \cap X \neq \emptyset} m(A), \ \forall X \in \mathcal{P}(U). \tag{8}$$

Belief and plausibility functions based on the same belief structure are connected by the dual property

$$\mathrm{Pl}(X) = 1 - \mathrm{Bel}(\sim X), \ \forall X \in \mathcal{P}(U), \tag{9}$$

and furthermore,

$$\mathrm{Bel}(X) \leq \mathrm{Pl}(X), \ \forall X \in \mathcal{P}(U). \tag{10}$$

There are strong connections between rough set theory and the Dempster–Shafer theory of evidence [35–37,44,55]. The following theorem shows that probabilities of lower and upper approximations are a dual pair of belief and plausibility functions [44,55].

Theorem 1. *Let (U, R, P) be a probabilistic approximation space, for any $X \subseteq U$, denote*

$$\mathrm{Bel}(X) = P(\underline{R}(X)), \qquad \mathrm{Pl}(X) = P(\overline{R}(X)). \tag{11}$$

Then Bel *and* Pl *are a dual pair of belief and plausibility functions on U respectively, and the corresponding basic probability assignment is*

$$m(Y) = \begin{cases} P(Y), & \text{if } Y \in U/R, \\ 0, & \text{otherwise}. \end{cases}$$

2.3. Information tables, decision tables, decision rules

The notion of information tables (sometimes called, information systems, data tables, attribute-value systems, knowledge representation systems etc.) provides a convenient tool for the representation of objects in terms of their attribute values [24,60].

Definition 5. An information table is a 2-tuple (U, A), where $U = \{x_1, x_2, \ldots, x_n\}$ is a non-empty, finite set of objects called the universe of discourse and $A = \{a_1, a_2, \ldots, a_m\}$ is a non-empty, finite set of attributes, such that $a : U \to V_a$ for any $a \in A$, i.e. $a(x) \in V_a$, $x \in U$, where $V_a = \{a(x) | x \in U\}$ is called the domain of a.

A *decision table* (sometimes called *decision system*) is a 2-tuple $S = (U, C \cup \{d\})$ where (U, C) is an information table, and $d \notin C$ is a special attribute called the decision. In this case, C is called the conditional attribute set, d is a mapping $d : U \to V_d$ from the universe of discourse U into the value set V_d, we assume, without any loss of generality, that $V_d = \{1, 2, \ldots, r\}$. Define

$$R_d = \{(x, y) \in U \times U | d(x) = d(y)\}. \tag{12}$$

Then we obtain a partition $U/R_d = \{D_1, D_2, \ldots, D_r\}$ of U into decision classes, where $D_j = \{x \in U | d(x) = j\}, j = 1, 2, \ldots, r$.

For any $B \subseteq C$, denote an equivalence relation (also called indiscernibility relation) R_B as

$$R_B = \bigcap_{a \in B} R_a = \{(x, y) \in U \times U | a(x) = a(y), \forall a \in B\}. \tag{13}$$

Since R_B is an equivalence relation on U, it forms a partition $U/R_B = \{[x]_B | x \in U\}$ of U, where $[x]_B$ denotes the equivalence class determined by x w.r.t. B, i.e., $[x]_B = \{y \in U | (x, y) \in R_B\}$.

If $R_C \subseteq R_d$, then the decision table $S = (U, C \cup \{d\})$ is referred to as *consistent*, it is said to be *inconsistent* otherwise.

For any $B \subseteq C$, define

$$\partial_B(x) = \{d(y) | y \in [x]_B\}, \; x \in U, \tag{14}$$

$\partial_B(x)$ is referred to as the *generalized decision value* of x w.r.t. B in $(U, C \cup \{d\})$ [12] and ∂_B is called the *generalized decision function* w.r.t. B in $(U, C \cup \{d\})$. It is straightforward that the decision table $(U, C \cup \{d\})$ is consistent if and only if $|\partial_C(x)| = 1$ for all $x \in U$, and it is inconsistent otherwise.

In the discussion to follow, the symbols $\wedge$ and $\vee$ denote the logical connectives “and" (conjunction) and “or"(disjunction), respectively. Any attribute-value pair (a, v), $v \in V_a$, $a \in B$, $B \subseteq C$, is called a *B-atomic property*. Any B-atomic property or conjunction of different B-atomic properties is called a *B-descriptor*. Let t be a B-descriptor, the attribute set occurring in t is denoted by $B(t)$. If $B(t) = B$, then t is called a *full B-descriptor*. Denote

$$\mathrm{FDES}(B) = \{t | t \text{ is a full } B\text{-descriptor}\}. \tag{15}$$

If (a, v) is an atomic property occurring in t, we simply say that $(a, v) \in t$.

The set of objects having descriptor t is called the *support* of t and is denoted by $\|t\|$, i.e., $\|t\| = \{x \in U | v = a(x), \forall (a, v) \in t\}$. If t and s are two atomic properties, then it can be observed that $\|t \wedge s\| = \|t\| \cap \|s\|$ and $\|t \vee s\| = \|t\| \cup \|s\|$. Clearly, if $C(t) = B$ and $x \in \|t\|$, then $[x]_B = \|t\|$.

For any B-descriptor t, let us define a function $\partial : U \to \mathcal{P}(V_d)$ as follows:

$$\partial(t) = \{d(y) | y \in \|t\|\}, \tag{16}$$

which is called the *generalized decision* of t in S. Any $(d, j), j \in \partial(t)$, is referred to as a *generalized decision descriptor* of t.

Let $X \subseteq U$ and $B \subseteq C$. The lower and upper approximations of X w.r.t. B, denoted by $\underline{R_B}(X)$ and $\overline{R_B}(X)$ respectively, are defined as follows:

$$\begin{aligned} \underline{R_B}(X) &= \cup\{\|t\| \mid \|t\| \subseteq X, t \in \mathrm{FDES}(B)\}, \\ \overline{R_B}(X) &= \cup\{\|t\| \mid \|t\| \cap X \neq \emptyset, t \in \mathrm{FDES}(B)\}. \end{aligned} \tag{17}$$

Clearly,

$$\begin{aligned} \underline{R_B}(X) &= \cup\{[x]_B \mid [x]_B \subseteq X\} = \{x \in U \mid [x]_B \subseteq X\}, \\ \overline{R_B}(X) &= \cup\{[x]_B \mid [x]_B \cap X \neq \emptyset\} = \{x \in U \mid [x]_B \cap X \neq \emptyset\}. \end{aligned} \tag{18}$$

Evidently, $\underline{R_B}(X) \subseteq \overline{R_B}(X)$. Elements in $\underline{R_B}(X)$ can be classified as members of X with complete certainty using attribute set B, whereas elements in $\overline{R_B}(X)$ can be classified as members of X with only partial certainty using attribute B. The class $\overline{R_B}(X) - \underline{R_B}(X)$ is referred to as *boundary* of X w.r.t B and is denoted by $\mathrm{BN}_B(X)$.

Let $j \in V_d$ and $t \in \mathrm{FDES}(B)$. If $\|t\| \subseteq \underline{R_B}(\|(d, j)\|)$ (and, respectively, $\|t\| \subseteq \overline{R_B}(\|(d, j)\|)$), then we call t a *lower (and, respectively, upper) approximation B-descriptor* of (d, j). The set of all lower (and, respectively, upper) approximation B-descriptors of (d, j) is denoted by $\underline{R_B}((d, j))$ (and, respectively, $\overline{R_B}((d, j))$). And also, if $\|t\| \subseteq \mathrm{BN}_B(\|(d, j)\|)$, then t is referred to as a *boundary descriptor* of (d, j) w.r.t. B. The set of all boundary descriptors of (d, j) w.r.t. B is denoted by $\mathrm{BNDES}_B((d, j))$.

Proposition 1 below shows that the approximations of decision classes can be expressed by means of the generalized decision.

Proposition 1. *Let $S = (U, C \cup \{d\})$ be a decision table. If $j \in V_d$, t is a C-descriptor, and $B \subseteq C$, then*

(1) $\underline{R_B}(\|(d, j)\|) = \cup\{\|t\| | t \in \mathrm{FDES}(B), \partial(t) = \{j\}\}$,
(2) $\overline{R_B}(\|(d, j)\|) = \cup\{\|t\| | t \in \mathrm{FDES}(B), j \in \partial(t)\}$,
(3) $\underline{R_B}((d, j)) = \{ t \in \mathrm{FDES}(B) | \partial(t) = \{j\}\}$,
(4) $\overline{R_B}((d, j)) = \{t \in \mathrm{FDES}(B) | j \in \partial(t)\}$.

The knowledge hidden in a decision table $S = (U, C \cup \{d\})$ may be discovered and expressed in the form of *decision rules*: $t \to s$, where $t = \wedge(a, v)$, $a \in B \subseteq C$, and $s = (d, j)$, $j \in V_d$, t and s are, respectively, called the *condition* and *decision* parts of the rule. We will say that an object $x \in U$ *supports a rule* $t \to s$ in the decision table S if and only if $x \in \|t\| \cap \|s\|$.

A decision rule $t \to (d, j)$ is referred to as *certain* in S if and only if $\|t\| \neq \emptyset$ and $\|t\| \subseteq \|(d, j)\|$, in such case, we denote $t \Rightarrow (d, j)$ instead of $t \to (d, j)$. A decision rule $t \to (d, j)$ is referred to as *an association rule* in S if and only if $\|t\| \cap \|(d, j)\| \neq \emptyset$. A decision rule $t \to (d, j)$ is referred to as *a possible rule* in S if and only if it is not certain, but $\|t\| \cap \|(d, j)\| \neq \emptyset$.

With each decision rule $t \to s = (d, j)$ in a decision table S, we associate a quantitative measure, called the *certainty*, of the rule in S and is defined by [25]:

$$Cer(t \to s) = \frac{|\ \|t\| \cap \|s\|\ |}{|\ \|t\|\ |}. \tag{19}$$

The quantity $Cer(t \to s)$ shows the degree to which objects supporting descriptor t also support decision s in S. If $Cer(t \to s) = \alpha$, then $(100\alpha)\%$ of objects supporting t also support s in S.

The following proposition shows that the types of decision rules can be expressed by means of the certainty factors of the rules as well as the lower and the upper approximations of each decision class w.r.t. the set of conditional attributes in a decision table.

Proposition 2. *Let $S = (U, C \cup \{d\})$ be a decision table. If $j \in V_d$, t is a C-descriptor, and $s = (d, j)$, then the decision rule $t \to s$*

(1) *is certain in S*
$$\iff \|t\| \subseteq \underline{R_{C(t)}}(\|(d, j)\|)$$
$$\iff t \in \underline{R_{C(t)}}((d, j))$$
$$\iff \partial(t) = \{j\}$$
$$\iff Cer(t \to (d, j)) = 1;$$

(2) *is an association rule in S*
$$\iff \|t\| \subseteq \overline{R_{C(t)}}(\|(d, j)\|)$$
$$\iff t \in \overline{R_{C(t)}}((d, j))$$
$$\iff j \in \partial(t)$$
$$\iff 0 < Cer(t \to (d, j)) \leq 1;$$

(3) *is a possible rule in S*
$$\iff \|t\| \subseteq \mathrm{BN}_{C(t)}(\|(d, j)\|)$$
$$\iff t \in \mathrm{BNDES}_{C(t)}((d, j))$$
$$\iff j \in \partial(t) \text{ and } |\ \partial(t)\ | \geq 2$$
$$\iff 0 < Cer(t \to (d, j)) < 1.$$

One can acquire certainty decision rules from consistent decision tables and uncertainty decision rules from inconsistent decision tables. In fact, if $|\ \partial_C(x)\ | = 1$, then the decision rule corresponding to (or supported by the objects in) the class $[x]_C$ is certain, otherwise, $|\ \partial_C(x)\ | \geq 2$, the decision rule corresponding to the class $[x]_C$ is uncertain.

A decision rule with too long a description means high prediction cost. To acquire concise decision rules from decision tables, knowledge reduction is needed. It is well-known that not all conditional attributes are necessary to depict the decision attributes before decision rules are generated. Thus knowledge reduction by reducing attributes is one of the main problems in the study of rough set theory (see e.g., [7,9,16,23,24,34,38,41,43,45,60,61,63]).

3. Multi-scale decision tables

In this section, we introduce the concept of multi-scale decision tables from the perspective of granular computation.

3.1. Multi-scale information tables

In a Pawak information table, each object can only take on one value under each attribute. However, in some real-life applications, one has to make decision with different levels of granulations. That is, an object may take on different values under the same attribute, depending on at which scale it is measured. In [42], we introduced a new concept called multi-scale information table from the perspective of granular computation which has different levels of granulations.

Definition 6. A *multi-scale information table* is a tuple $S = (U, A)$, where

- $U = \{x_1, x_2, \dots, x_n\}$ is a non-empty, finite set of objects called the universe of discourse;
- $A = \{a_1, a_2, \dots, a_m\}$ is a non-empty, finite set of attributes, and each $a_j \in A$ is a multi-scale attribute, i.e., for the same object in U, attribute a_j can take on different values at different scales.

In the discussion to follow, we always assume that all the attributes have the same number I of levels of granulations. Hence, a multi-scale information table can be represented as a table $(U, \{a_j^k | k = 1, 2, \dots, I, j = 1, 2, \dots, m\})$, where $a_j^k : U \to V_j^k$ is a surjective function and V_j^k is the domain of the kth scale attribute a_j^k. For $1 \le k \le I - 1$, there exists a surjective function $g_j^{k,k+1} : V_j^k \to V_j^{k+1}$ such that $a_j^{k+1} = g_j^{k,k+1} \circ a_j^k$, i.e.

$$a_j^{k+1}(x) = g_j^{k,k+1}\left(a_j^k(x)\right), \quad x \in U, \tag{20}$$

where $g_j^{k,k+1}$ is called a *granular information transformation function*.

Definition 7. Let U be a nonempty set, and $\mathcal{A}_1$ and $\mathcal{A}_2$ be two partitions of U. If for each $A_1 \in \mathcal{A}_1$, there exists $A_2 \in \mathcal{A}_2$ such that $A_1 \subseteq A_2$, then we say that $\mathcal{A}_1$ is finer than $\mathcal{A}_2$ or $\mathcal{A}_2$ is coarser than $\mathcal{A}_1$, and is denoted as $\mathcal{A}_1 \sqsubseteq \mathcal{A}_2$. Furthermore, if there exist $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$ such that $A_1 \subset A_2$, then we say that $\mathcal{A}_1$ is strictly finer than $\mathcal{A}_2$, and is denoted as $\mathcal{A}_1 \sqsubset \mathcal{A}_2$.

For $k \in \{1, 2, \dots, I\}$, we denote $A^k = \left\{a_j^k | j = 1, 2, \dots, m\right\}$. Then a multi-scale information table $S = (U, A)$ can be decomposed into I information tables $S^k = (U, A^k)$, $k = 1, 2, \dots, I$. According to [42], we can conclude the following:

Proposition 3. *Let* $S = (U, A) = \left(U, \left\{a_j^k | k = 1, 2, \dots, I, j = 1, 2, \dots, m\right\}\right)$ *be a multi-scale information table, and* $B \subseteq A$, *for* $k \in \{1, 2, \dots, I\}$, *denote*

$$\begin{aligned} &R_{B^k} = \{(x, y) \in U \times U | a^k(x) = a^k(y),\ \forall a \in B\}, \\ &[x]_{B^k} = \{y \in U | (x, y) \in R_{B^k}\} = \{y \in U | a^k(x) = a^k(y),\ \forall a \in B\}. \\ &U/R_{B^k} = \{[x]_{B^k} | x \in U\}. \end{aligned} \tag{21}$$

Then

$$\begin{aligned} &R_{B^1} \subseteq R_{B^2} \subseteq \cdots \subseteq R_{B^I}, \\ &[x]_{B^1} \subseteq [x]_{B^2} \subseteq \cdots \subseteq [x]_{B^I},\ x \in U, \\ &U/R_{B^1} \sqsubseteq U/R_{B^2} \sqsubseteq \cdots \sqsubseteq U/R_{B^I}. \end{aligned} \tag{22}$$

For $B \subseteq A$ and $X \subseteq U$, since $U/R_{B^1} \sqsubseteq U/R_{B^2} \sqsubseteq \cdots \sqsubseteq U/R_{B^I}$, according to Yao [48], we can obtain a nested sequence of rough set approximations as follows:

$$\begin{aligned} &\underline{R_{B^I}}(X) \subseteq \underline{R_{B^{I-1}}}(X) \subseteq \cdots \subseteq \underline{R_{B^2}}(X) \subseteq \underline{R_{B^1}}(X) \subseteq X, \\ &X \subseteq \overline{R_{B^1}}(X) \subseteq \overline{R_{B^2}}(X) \subseteq \cdots \subseteq \overline{R_{B^{I-1}}}(X) \subseteq \overline{R_{B^I}}(X). \end{aligned} \tag{23}$$

Therefore, we can have nested sequences of the positive, the boundary, the positive, and the negative regions:

$$\begin{aligned} &\mathrm{POS}_{B^I}(X) \subseteq \mathrm{POS}_{B^{I-1}}(X) \subseteq \cdots \subseteq \mathrm{POS}_{B^2}(X) \subseteq \mathrm{POS}_{B^1}(X), \\ &\mathrm{BN}_{B^1}(X) \subseteq \mathrm{BN}_{B^2}(X) \subseteq \cdots \subseteq \mathrm{BN}_{B^{I-1}}(X) \subseteq \mathrm{BN}_{B^I}(X), \\ &\mathrm{NEG}_{B^I}(X) \subseteq \mathrm{NEG}_{B^{I-1}}(X) \subseteq \cdots \subseteq \mathrm{NEG}_{B^2}(X) \subseteq \mathrm{NEG}_{B^1}(X), \end{aligned} \tag{24}$$

Consequently, we obtain a sequence of accuracies for approximations w.r.t. different scales:

$$\alpha_{B^I}(X) \le \alpha_{B^{I-1}}(X) \le \cdots \le \alpha_{B^2}(X) \le \alpha_{B^1}(X). \tag{25}$$

By employing Theorem 1 and Eq. (23) we can conclude following

Proposition 4. *Let* $S = (U, A) = \left(U, \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\}\right)$ *be a multi-scale information table, and* $B \subseteq A$, *for* $k \in \{1, 2, \ldots, I\}$, *denote*

$$\begin{aligned} \mathrm{Bel}_{B^k}(X) &= P(\underline{R_{B^k}}(X)) = \frac{|\underline{R_{B^k}}(X)|}{|U|}, \\ \mathrm{Pl}_{B^k}(X) &= P(\overline{R_{B^k}}(X)) = \frac{|\overline{R_{B^k}}(X)|}{|U|}. \end{aligned} \tag{26}$$

Then $\mathrm{Bel}_{B^k} : \mathcal{P}(U) \to [0, 1]$ *and* $\mathrm{Pl}_{B^k} : \mathcal{P}(U) \to [0, 1]$ *are a dual pair of belief and plausibility functions on U, and the corresponding basic probability assignment* $m_{B^k} : \mathcal{P}(U) \to [0, 1]$ *is*

$$m_{B^k}(Y) = \begin{cases} P(Y) = \frac{|Y|}{|U|}, & \text{if } Y \in U/R_{B^k}, \\ 0, & \text{otherwise}. \end{cases} \tag{27}$$

Moreover, the belief and plausibility functions satisfy the following properties:

(1) $\mathrm{Bel}_{B^I}(X) \le \mathrm{Bel}_{B^{I-1}}(X) \le \cdots \le \mathrm{Bel}_{B^2}(X) \le \mathrm{Bel}_{B^1}(X) \le P(X)$,
(2) $P(X) \le \mathrm{Pl}_{B^1}(X) \le \mathrm{Pl}_{B^2}(X) \le \cdots \le \mathrm{Pl}_{B^{I-1}}(X) \le \mathrm{Pl}_{B^I}(X)$,
(3) $B \subseteq C \subseteq A \Longrightarrow \mathrm{Bel}_{B^k}(X) \le \mathrm{Bel}_{C^k}(X) \le P(X) \le \mathrm{Pl}_{C^k}(X) \le \mathrm{Pl}_{B^k}(X)$.

Definition 8. A system $S = (U, C \cup \{d\}) = \left(U, \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ is referred to as a *multi-scale decision table*, where $(U, C) = \left(U, \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\}\right)$ is a multi-scale information table and $d \notin \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\}$, $d : U \to V_d$, is a special attribute called the decision.

According to Definition 8, a multi-scale decision table $S = (U, C \cup \{d\}) = \left(U, \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ can be decomposed into I decision tables $S^k = \left(U, \left\{a_j^k | j = 1, 2, \ldots, m\right\} \cup \{d\}\right) = (U, C^k \cup \{d\})$, $\left(C^k = \left\{a_j^k | j = 1, 2, \ldots, m\right\}\right)$, $k = 1, 2, \ldots, I$, with the same decision d.

Table 1
A multi-scale decision table with three levels of granulations.

U	a_1^1	a_1^2	a_1^3	a_2^1	a_2^2	a_2^3	a_3^1	a_3^2	a_3^3	a_4^1	a_4^2	a_4^3	d
x_1	1	E	Y	1	E	Y	1	S	Y	1	S	Y	1
x_2	2	G	Y	2	E	Y	1	S	Y	1	S	Y	1
x_3	3	G	Y	3	G	Y	2	S	Y	2	S	Y	1
x_4	4	F	N	4	F	N	3	M	N	3	M	N	1
x_5	5	B	N	5	F	N	4	L	N	4	L	N	1
x_6	6	B	N	6	B	N	5	L	N	4	L	N	1
x_7	4	F	N	4	F	N	1	S	Y	1	S	Y	2
x_8	5	B	N	5	F	N	1	S	Y	1	S	Y	2
x_9	6	B	N	6	B	N	2	S	Y	2	S	Y	2
x_{10}	4	F	N	4	F	N	3	M	N	1	S	Y	1
x_{11}	5	B	N	5	F	N	4	L	N	1	S	Y	1
x_{12}	6	B	N	6	B	N	5	L	N	2	S	Y	1

Table 2
The decision table with the first level of granulation of Table 1.

U	a_1^1	a_2^1	a_3^1	a_4^1	d
x_1	1	1	1	1	1
x_2	2	2	1	1	1
x_3	3	3	2	2	1
x_4	4	4	3	3	1
x_5	5	5	4	4	1
x_6	6	6	5	4	1
x_7	4	4	1	1	2
x_8	5	5	1	1	2
x_9	6	6	2	2	2
x_{10}	4	4	3	1	1
x_{11}	5	5	4	1	1
x_{12}	6	6	5	2	1

Table 3
The decision table with the second level of granulation of Table 1.

U	a_1^2	a_2^2	a_3^2	a_4^2	d
x_1	E	E	S	S	1
x_2	G	E	S	S	1
x_3	G	G	S	S	1
x_4	F	F	M	M	1
x_5	B	F	L	L	1
x_6	B	B	L	L	1
x_7	F	F	S	S	2
x_8	B	F	S	S	2
x_9	B	B	S	S	2
x_{10}	F	F	M	S	1
x_{11}	B	F	L	S	1
x_{12}	B	B	L	S	1

Table 4
The decision table with the third level of granulation of Table 1.

U	a_1^3	a_2^3	a_3^3	a_4^3	d
x_1	Y	Y	Y	Y	1
x_2	Y	Y	Y	Y	1
x_3	Y	Y	Y	Y	1
x_4	N	N	N	N	1
x_5	N	N	N	N	1
x_6	N	N	N	N	1
x_7	N	N	Y	Y	2
x_8	N	N	Y	Y	2
x_9	N	N	Y	Y	2
x_{10}	N	N	N	Y	1
x_{11}	N	N	N	Y	1
x_{12}	N	N	N	Y	1

Definition 9. A multi-scale decision table S is referred to as consistent if the decision table under the first (finest) level of scale, $S^1 = \left(U, \left\{a_j^1 | j = 1, 2, \ldots, m\right\} \cup \{d\}\right) = (U, C^1 \cup \{d\})$, is consistent, and S is called inconsistent if S^1 is an inconsistent decision table.

Example 1. Table 1 is an example of a multi-scale decision table $\left(U, \left\{a_j^k | k = 1, 2, 3, j = 1, 2, 3, 4\right\} \cup \{d\}\right)$, where $U = \{x_1, x_2, \ldots, x_{12}\}$, $C = \{a_1, a_2, a_3, a_4\}$. The table has three levels of granulations, where "E", "G", "F", "B", "S", "M", "L", "Y", and "N" stand for, respectively, "Excellent", "Good", "Fair", "Bad", "Small", "Medium", "Large", "Yes", and "No". For these levels of granularities, the system is associated with three decision tables which are described as Tables 2–4, respectively.

4. Optimal scale selection in multi-scale decision tables

Knowledge acquisition in the sense of rule induction from a multi-scale decision table is an important issue. As we know from last section, a multi-scale decision table having I levels of granulations can be decomposed into I decision tables, however, not all decision tables are consistent with some requirements to the decision table under the first (finest) level of scale. So, it is critical to select the optimal level of details corresponding a suitable decision table before decision rules are produced. In this section, we investigate optimal scale selection with different requirements in multi-scale decision tables.

4.1. Optimal scale selection in consistent multi-scale decision tables

For a consistent multi-scale decision table $S = (U, C \cup \{d\}) = \left(U, \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$, we have $R_{C^1} \subseteq R_d$. For $1 \le i < k \le I$, if $S^k = \left(U, \left\{a_j^k | j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ is a consistent decision table, i.e. $R_{C^k} \subseteq R_d$, then, by Proposition 3, we can observe that $R_{C^i} \subseteq R_{C^k} \subseteq R_d$. Hence, $\left(U, \left\{a_j^i | j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ is also a consistent decision table.

Definition 10. Let $S = (U, C \cup \{d\}) = \left(U, \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ be a consistent multi-scale decision table. The kth level of scale is said to be optimal if S^k is consistent and S^{k+1} (if there exists $k + 1$) is inconsistent.

According to Definition 10, we can see that the optimal scale of a consistent multi-scale decision table is the best scale for decision making or classification in the multi-scale decision table. And k is the optimal scale if and only if k is the maximal number such that S^k is a consistent decision table.

Theorem 2. *Let* $S = (U, C \cup \{d\}) = \left(U, \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ *be a consistent multi-scale decision table which has I levels of granulations. For* $k \in \{1, 2, \ldots, I\}$*, then the following statements are equivalent:*

(1) $S^k = \left(U, \left\{a_j^k | j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ *is a consistent decision table, i.e.,* $R_{C^k} \subseteq R_d$*,*
(2) $\sum_{j=1}^r \mathrm{Bel}_{C^k}(D_j) = 1$,
(3) $\sum_{j=1}^r \mathrm{Pl}_{C^k}(D_j) = 1$.

Proof

"(1) $\Rightarrow$ (2)" For any $j \in \{1, 2, \ldots, r\}$, denote

$$\mathcal{J}_{C^k}(D_j) = \{[y]_{C^k} \in U/R_{C^k} | [y]_{C^k} \subseteq D_j\}.$$

Since $R_{C^k} \subseteq R_d$, we see that $\mathcal{J}_{C^k}(D_j)$ forms a partition of D_j. Then we have

$$\begin{aligned} \mathrm{Bel}_{C^k}(D_j) &= \sum\{m_{c^k}(X) | X \subseteq D_j\} = \sum\{m_{c^k}([x]_{C^k}) | [x]_{C^k} \in U/R_{C^k}, [x]_{C^k} \subseteq D_j\} \\ &= \sum\{m_{c^k}([x]_{C^k}) | [x]_{C^k} \in \mathcal{J}_{C^k}(D_j)\} = \sum\left\{P([x]_{C^k}) | [x]_{C^k} \in \mathcal{J}_{C^k}(D_j)\right\} \\ &= P(D_j). \end{aligned}$$

It follows that

$$\sum_{j=1}^r \mathrm{Bel}_{C^k}(D_j) = \sum_{j=1}^r P(D_j) = 1.$$

"(2) $\Rightarrow$ (1)" Assume that $\sum\limits_{j=1}^r \mathrm{Bel}_{C^k}(D_j) = 1$. Define

$$\mathcal{J}_{C^1}([x]_{C^k}) = \left\{[y]_{C^1} \in U/R_{C^1} | [y]_{C^1} \subseteq [x]_{C^k}\right\}, \quad x \in U.$$

It is easy to see that $\mathcal{J}_{C^1}([x]_{C^k})$ forms a partition of $[x]_{C^k}$. Then, for any $j \in \{1, 2, \ldots, r\}$, we have

$$\begin{aligned} \mathrm{Bel}_{C^k}(D_j) &= \sum\left\{m_{c^k}([x]_{C^k}) | [x]_{C^k} \in U/R_{C^k}, [x]_{C^k} \subseteq D_j\right\} \\ &= \sum\left\{P([x]_{C^k}) | [x]_{C^k} \in U/R_{C^k}, [x]_{C^k} \subseteq D_j\right\} \\ &= \sum\left\{P([y]_{C^1}) | [y]_{C^1} \in \mathcal{J}_{C^1}([x]_{C^k}), [x]_{C^k} \in U/R_{C^k}, [x]_{C^k} \subseteq D_j\right\} \\ &\leq \sum\left\{P([x]_{C^1}) | [x]_{C^1} \subseteq D_j\right\} = P(D_j). \end{aligned}$$

Since $1 = \sum_{j=1}^r \mathrm{Bel}_{C^k}(D_j) \leq \sum_{j=1}^r P(D_j) = 1$, by Proposition 4, we can conclude that

$$\mathrm{Bel}_{C^k}(D_j) = P(D_j), \quad \forall j \in \{1, 2, \ldots, r\}.$$

From which we can see that $\left\{[x]_{C^k} | [x]_{C^k} \in U/R_{C^k}, [x]_{C^k} \subseteq D_j\right\}$ forms a partition of D_j. Since $\{D_j | j \in \{1, 2, \ldots, r\}\}$ is a partition of U, we conclude that $\left\{[x]_{C^k} | [x]_{C^k} \subseteq D_j, j \in \{1, 2, \ldots, r\}\right\}$ forms a partition of U. Hence for any $x \in U$, there exists $j \in \{1, 2, \ldots, r\}$ such that $[x]_{C^k} \subseteq D_j$. Evidently,

$$x \in [x]_{C^k} \subseteq D_j \Longleftrightarrow [x]_d = D_j.$$

Thus $[x]_{C^k} \subseteq [x]_d$ for all $x \in U$, that is, $R_{C^k} \subseteq R_d$.

"(1) $\Rightarrow$ (3)" Since $R_{C^k} \subseteq R_d$, we have $[x]_{C^k} \subseteq [x]_d$ for all $x \in U$. Define

$$\mathcal{J}_{C^k}(D_j) = \{[x]_{C^k} \in U/R_{C^k} | [x]_{C^k} \subseteq D_j\}, \quad j \in \{1, 2, \ldots, r\}.$$

It is easy to see from $R_{C^k} \subseteq R_d$ that $\mathcal{J}_{C^k}(D_j)$ forms a partition of D_j, and moreover,

$$[x]_{C^k} \cap D_j \neq \emptyset \Longleftrightarrow [x]_{C^k} \subseteq D_j, \quad \forall x \in U.$$

Hence

$$\begin{aligned}\mathrm{Pl}_{C^k}(D_j) &= \sum\left\{m_{C^k}(Y)|Y\cap D_j\neq\emptyset\right\}\\ &= \sum\left\{m_{C^k}([x]_{C^k})|[x]_{C^k}\in U/R_{C^k},\,[x]_{C^k}\cap D_j\neq\emptyset\right\}\\ &= \sum\left\{m_{C^k}([x]_{C^k})|[x]_{C^k}\in U/R_{C^k},\,[x]_{C^k}\subseteq D_j\right\}\\ &= \sum\left\{P([x]_{C^k})|[x]_{C^k}\in \mathcal{J}_{C^k}(D_j)\right\}\\ &= P(D_j),\quad \forall j\in\{1,2,\ldots,r\}.\end{aligned}$$

It follows that

$$\sum_{j=1}^{r}\mathrm{Pl}_{C^k}(D_j)=\sum_{j=1}^{r}P(D_j)=1.$$

"(3) ⇒ (1)" Assume that $\sum_{j=1}^{r}\mathrm{Pl}_{C^k}(D_j)=1$. Since S is consistent, we have $1=\sum_{j=1}^{r}\mathrm{Pl}_{C^1}(D_j)\geq\sum_{j=1}^{r}\mathrm{Pl}_{C^k}(D_j)=1$. Then by Proposition 4 we have $\mathrm{Pl}_{C^k}(D_j)=\mathrm{Pl}_{C^1}(D_j)=P(D_j)$ for all $j\in\{1,2,\ldots,r\}$, that is,

$$P\left(\overline{R_{C^k}}(D_j)\right)=P\left(\overline{R_{C^1}}(D_j)\right)=P(D_j).$$

By Eq. (23), we observe that $\overline{R_{C^k}}(D_j)\supseteq\overline{R_{C^1}}(D_j)\supseteq D_j$, then we conclude that $\overline{R_{C^k}}(D_j)=\overline{R_{C^1}}(D_j)=D_j$ for all $j\in\{1,2,\ldots,r\}$. Thus

$$\overline{R_{C^k}}([x]_d)=\overline{R_{C^1}}([x]_d)=[x]_d,\qquad\forall x\in U.$$

Given $x\in U$ and for any $y\in[x]_{C^k}$, notice that $[y]_{C^k}=[x]_{C^k}$, then $[y]_{C^k}\cap[x]_d=[x]_{C^k}\cap[x]_d\neq\emptyset$, that is, $y\in\overline{R_{C^k}}([x]_d)=[x]_d$, and in turn, $[x]_{C^k}\subseteq[x]_d$. It follows that $R_{C^k}\subseteq R_d$. □

In terms of Theorem 2 and Proposition 4, we can conclude following

Theorem 3. *Let* $S=(U,C\cup\{d\})=\left(U,\left\{a_j^k|k=1,2,\ldots,I,j=1,2,\ldots,m\right\}\cup\{d\}\right)$ *be a consistent multi-scale decision table which has I levels of granulations. For* $k\in\{1,2,\ldots,I\}$*, then the following statements are equivalent:*

(1) *the kth level of scale is the optimal scale.*
(2)

$$\sum_{j=1}^{r}\mathrm{Bel}_{C^k}(D_j)=1. \tag{28}$$

And (if there is $k+1\leq I$*)*

$$\sum_{j=1}^{r}\mathrm{Bel}_{C^{k+1}}(D_j)<1. \tag{29}$$

(3)

$$\sum_{j=1}^{r}\mathrm{Pl}_{C^k}(D_j)=1. \tag{30}$$

And (if there is $k+1\leq I$*)*

$$\sum_{j=1}^{r}\mathrm{Pl}_{C^{k+1}}(D_j)>1. \tag{31}$$

Theorem 3 shows that, in a consistent multi-scale decision table, the kth level of scale is the optimal scale if and only if k is the maximum number such that the sum of degrees of belief (as well as the degrees of plausibility) of all decision classes in S^k is 1.

After we select the optimal scale k, for making decision, we can obtain the classification rule set based on the computing reducts of the kth decision table S^k, because this issue is not the main objective of this paper, we will not discuss it here, and for the detail we refer the readers to [42].

4.2. Optimal scale selection in inconsistent multi-scale decision tables

Let $S = (U, C \cup \{d\}) = \left(U, \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ be a multi-scale decision table which has I levels of granulations. For $1 \le i < k \le I$, if $\left(U, \left\{a_j^i | j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ is an inconsistent decision table, then it can easily be observed that $\left(U, \left\{a_j^k | j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ is also an inconsistent decision table.

For $k \in \{1, 2, \ldots, I\}$, and $X \subseteq U$, define

$$\begin{aligned} \underline{R_{C^k}}(X) &= \{x \in U | [x]_{C^k} \subseteq X\} = \{x \in U | P(X|[x]_{C^k}) = 1\}, \\ \overline{R_{C^k}}(X) &= \{x \in U | [x]_{C^k} \cap X \neq \emptyset\} = \{x \in U | P(X|[x]_{C^k}) > 0\}, \end{aligned} \tag{32}$$

where $R_{C^k} = \{(x, y) \in U \times U | a^k(x) = a^k(y), \forall a \in C\}$ and $[x]_{C^k} = \{y \in U | (x, y) \in R_{C^k}\}$.

We denote

$$\begin{aligned} L_{C^k}(d) &= \left(\underline{R_{C^k}}(D_1), \underline{R_{C^k}}(D_2), \ldots, \underline{R_{C^k}}(D_r)\right), \\ H_{C^k}(d) &= \left(\overline{R_{C^k}}(D_1), \overline{R_{C^k}}(D_2), \ldots, \overline{R_{C^k}}(D_r)\right), \\ \mu_{C^k}(x) &= \left(P(D_1|[x]_{C^k}), P(D_2|[x]_{C^k}), \ldots, P(D_r|[x]_{C^k})\right), \quad x \in U, \\ \gamma_{C^k}(x) &= \{D_{j_i} \in U/R_d | P(D_{j_i}|[x]_{C^k}) = \max_{1 \le j \le r} P(D_j|[x]_{C^k})\}, \quad x \in U, \\ \mathrm{Bel}_{C^k}(d) &= \left(\mathrm{Bel}_{C^k}(D_1), \mathrm{Bel}_{C^k}(D_2), \ldots, \mathrm{Bel}_{C^k}(D_r)\right), \\ \mathrm{Pl}_{C^k}(d) &= \left(\mathrm{Pl}_{C^k}(D_1), \mathrm{Pl}_{C^k}(D_2), \ldots, \mathrm{Pl}_{C^k}(D_r)\right), \\ \partial_{C^k}(x) &= \{d(y) | y \in [x]_{C^k}\}, \quad x \in U, \end{aligned}$$

where $P(D_j|[x]_{C^k}) = \frac{|D_j \cap [x]_{C^k}|}{|[x]_{C^k}|}$, $\mathrm{Bel}_{C^k}(D_j) = P(\underline{R_{C^k}}(D_j)) = \frac{|\underline{R_{C^k}}(D_j)|}{|U|}$, and $\mathrm{Pl}_{C^k}(D_j) = P(\overline{R_{C^k}}(D_j)) = \frac{|\overline{R_{C^k}}(D_j)|}{|U|}$.

$L_{C^k}(d)$ and H_{C^k} are referred to as the lower approximation distribution and upper approximation distribution of decision classes U/R_d under the kth scale in S, respectively. $\mu_{C^k}(x)$ is called the probability distribution of decision classes U/R_d for object x under the kth scale in S, and $\gamma_{C^k}(x)$ is called the maximum distribution of decision classes U/R_d for object x under the kth scale in S. $\mathrm{Bel}_{C^k}(d)$ and $\mathrm{Pl}_{C^k}(d)$ are said to be the belief distribution and plausibility distribution of decision classes U/R_d under the kth scale in S, respectively. And $\partial_{C^k}(x)$ is the generalized decision values of object x under the kth scale in S. According to Proposition 3 it is easy to see that

$$\partial_{C^1}(x) \subseteq \partial_{C^2}(x) \subseteq \cdots \subseteq \partial_{C^{I-1}}(x) \subseteq \partial_{C^I}(x), \; x \in U. \tag{33}$$

Definition 11. Let $S = (U, C \cup \{d\}) = \left(U, \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ be an inconsistent multi-scale decision table which has I levels of granulations. For $k \in \{1, 2, \ldots, I\}$, we say that

(1) $S^k = (U, C^k \cup \{d\}) = \left(U, \left\{a_j^k | j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ is lower approximation consistent to S if $L_{C^k}(d) = L_{C^1}(d)$. And, the kth level of scale is said to be the lower approximation optimal scale of S if S^k is lower approximation consistent to S and S^{k+1} (if there is $k+1$) is not lower approximation consistent to S.
(2) $S^k = (U, C^k \cup \{d\}) = \left(U, \left\{a_j^k | j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ is upper approximation consistent to S if $H_{C^k}(d) = H_{C^1}(d)$. And, the kth level of scale is said to be the upper approximation optimal scale of S if S^k is upper approximation consistent to S and S^{k+1} (if there is $k+1$) is not upper approximation consistent to S.
(3) $S^k = (U, C^k \cup \{d\}) = \left(U, \left\{a_j^k | j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ is distribution consistent to S if $\mu_{C^k}(x) = \mu_{C^1}(x)$ for all $x \in U$. And, the kth level of scale is said to be the distribution optimal scale of S if S^k is distribution consistent to S and S^{k+1} (if there is $k+1$) is not distribution consistent to S.
(4) $S^k = (U, C^k \cup \{d\}) = \left(U, \left\{a_j^k | j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ is maximum distribution consistent to S if $\gamma_{C^k}(x) = \gamma_{C^1}(x)$ for all $x \in U$. And, the kth level of scale is said to be the maximum distribution optimal scale of S if S^k is maximum distribution consistent to S and S^{k+1} (if there is $k+1$) is not maximum distribution consistent to S.

(5) $S^k = (U, C^k \cup \{d\}) = \left(U, \left\{a_j^k | j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ is belief distribution consistent to S if $\mathrm{Bel}_{C^k}(d) = \mathrm{Bel}_{C^1}(d)$. And, the kth level of scale is said to be the belief distribution optimal scale of S if S^k is belief distribution consistent to S and S^{k+1} (if there is $k+1$) is not belief distribution consistent to S.

(6) $S^k = (U, C^k \cup \{d\}) = \left(U, \left\{a_j^k | j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ is plausibility distribution consistent to S if $\mathrm{Pl}_{C^k}(d) = \mathrm{Pl}_{C^1}(d)$. And, the kth level of scale is said to be the plausibility distribution optimal scale of S if S^k is plausibility distribution consistent to S and S^{k+1} (if there is $k+1$) is not plausibility distribution consistent to S.

(7) $S^k = (U, C^k \cup \{d\}) = \left(U, \left\{a_j^k | j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ is generalized decision consistent to S if $\partial_{C^k}(x) = \partial_{C^1}(x)$ for all $x \in U$. And, the kth level of scale is said to be the generalized decision optimal scale of S if S^k is generalized decision consistent to S and S^{k+1} (if there is $k+1$) is not generalized decision consistent to S.

In an inconsistent multi-scale decision table which has I levels of granulations, it can be observed that $S^k = \left(U, \left\{a_j^k | j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ is an inconsistent decision table for all $k \in \{1, 2, \ldots, I\}$. Moreover, we can see that

- S^k is lower approximation consistent to S if and only if S^k preserves the lower approximations of all decision classes of the finest scale decision table S^1, in this case, an object supports a certain decision rule derived from S^1 if and only if it supports a certain decision rule derived from S^k. And k is the lower approximation optimal scale of S if and only if k is the maximal number such that S^k preserves the lower approximations of all decision classes of S^1.
- S^k is upper approximation consistent to S if and only if S^k preserves the upper approximations of all decision classes of the finest scale decision table S^1, in this case, an object supports an association rule derived from S^1 if and only if it supports an association rule derived from S^k. And k is the upper approximation optimal scale of S if and only if k is the maximal number such that S^k preserves the upper approximations of all decision classes of S^1.
- S^k is distribution consistent to S if and only if S^k preserves the degree in which every object belongs to each decision class of the finest scale decision table S^1. And k is the distribution optimal scale of S if and only if k is the maximal number such that S^k preserves the degree in which every object belongs to each decision class of S^1.
- S^k is maximum distribution consistent to S if and only if S^k preserves all maximum confidence degree of decision rules of the finest scale decision table S^1. And k is the maximum distribution optimal scale of S if and only if k is the maximal number such that S^k preserves all maximum confidence degree of decision rules of S^1.
- S^k is belief distribution consistent to S if and only if S^k preserves the same belief degree of each decision class in the finest scale decision table S^1. And k is the belief distribution optimal scale of S if and only if k is the maximal number such that S^k preserves the same belief degree of each decision class in S^1.
- S^k is plausibility distribution consistent to S if and only if S^k preserves the same plausibility degree of each decision class in the finest scale decision table S^1. And k is the plausibility distribution optimal scale of S if and only if k is the maximal number such that S^k preserves the same plausibility degree of each decision class in S^1.
- S^k is generalized decision consistent to S if and only if S^k keeps the generalized decision values of the finest scale decision table S^1. And k is the generalized decision optimal scale of S if and only if k is the maximal number such that S^k keeps the generalized decision values of S^1.

It is important to clarify the interrelationships among the defined types of optimal scale in Definition 11.

Theorem 4. *Let* $S = (U, C \cup \{d\}) = \left(U, \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ *be an inconsistent multi-scale decision table which has I levels of granulations. For* $k \in \{1, 2, \ldots, I\}$, *if* $\mu_{C^k}(x) = \mu_{C^1}(x)$ *for all* $x \in U$, *then*

(1) $L_{C^k}(d) = L_{C^1}(d)$.
(2) $H_{C^k}(d) = H_{C^1}(d)$.
(3) $\gamma_{C^k}(x) = \gamma_{C^1}(x)$ *for all* $x \in U$.

Proof

(1) Notice that $R_{C^1} \subseteq R_{C^k}$, by Eq. (23), we have $\underline{R_{C^k}}(D_j) \subseteq \underline{R_{C^1}}(D_j)$ for all $D_j \in U/R_d$. On the other hand, for any $D_j \in U/R_d$ and $x \in U$, if $x \in \underline{R_{C^1}}(D_j)$, by the definition of lower approximation, we have $[x]_{C^1} \subseteq D_j$, then $P(D_j|[x]_{C^1}) = 1$. By the assumption we conclude that $P(D_j|[x]_{C^k}) = P(D_j|[x]_{C^1}) = 1$, thus $x \in \underline{R_{C^k}}(D_j)$, from which follows that $\underline{R_{C^k}}(D_j) = \underline{R_{C^1}}(D_j)$. Consequently, $L_{C^k}(d) = L_{C^1}(d)$.

(2) Since $R_{C^1} \subseteq R_{C^k}$, by Eq. (23), we have $\overline{R_{C^1}}(D_j) \subseteq \overline{R_{C^k}}(D_j)$ for all $D_j \in U/R_d$. On the other hand, for any $D_j \in U/R_d$ and $x \in U$, if $x \in \overline{R_{C^k}}(D_j)$, by the definition of upper approximation, we see that $P(D_j|[x]_{C^k}) > 0$. By the assumption we conclude that $P(D_j|[x]_{C^1}) = P(D_j|[x]_{C^k}) > 0$, thus $x \in \overline{R_{C^1}}(D_j)$, from which follows that $\overline{R_{C^k}}(D_j) \subseteq \overline{R_{C^1}}(D_j)$. Therefore, $\overline{R_{C^k}}(D_j) = \overline{R_{C^1}}(D_j)$ for all $D_j \in U/R_d$. Consequently, $H_{C^k}(d) = H_{C^1}(d)$.

(3) It is straightforward. □

Table 5
A multi-scale decision table.

U	a^1	a^2	d	∂_{a^1}	∂_{a^2}	μ_{a^1}	μ_{a^2}
x_1	1	S	1	{1}	{1, 2}	1	4/5
x_2	1	S	1	{1}	{1, 2}	1	4/5
x_3	2	S	1	{1, 2}	{1, 2}	2/3	4/5
x_4	2	S	1	{1, 2}	{1, 2}	2/3	4/5
x_5	2	S	2	{1, 2}	{1, 2}	1/3	1/5
x_6	3	L	2	{2}	{2}	1	1

Table 6
A multi-scale decision table.

U	a^1	a^2	d	∂_{a^1}	∂_{a^2}	μ_{a^1}	μ_{a^2}
x_1	1	S	1	{1, 2}	{1, 2}	1/2	2/5
x_2	1	S	2	{1, 2}	{1, 2}	1/2	3/5
x_3	2	S	2	{1, 2}	{1, 2}	2/3	3/5
x_4	2	S	1	{1, 2}	{1, 2}	1/3	2/5
x_5	2	S	2	{1, 2}	{1, 2}	2/3	3/5
x_6	3	L	2	{2}	{2}	1	1

The converse of Theorem 4 is not always true, to see following two examples.

Example 2. Table 5 gives an example of multi-scale decision table $S = (U, C \cup \{d\})$ which has 2 levels of granulations and one attribute a, where $U = \{x_1, x_2, \ldots, x_6\}$, $C = \{a\}$. We also list the generalized decision functions and the rough membership functions (decision distribution functions) for the two levels of granulations. It can easily be verified from Table 5 that $k = 1$ is the distribution optimal scale of S, it is also the lower approximation optimal scale and the upper approximation optimal scale. However, the maximum distribution optimal scale of S is 2. That is, in general, the maximum distribution optimal scale of S is not less than the distribution optimal scale of S.

Example 3. Table 6 gives another example of multi-scale decision table $S = (U, C \cup \{d\})$ which has 2 levels of granulations and one attribute a, where $U = \{x_1, x_2, \ldots, x_6\}$, $C = \{a\}$. It can be calculated that $L_{a^1}(d) = L_{a^2}(d) = (\emptyset, \{x_6\})$ and $H_{a^1}(d) = H_{a^2}(d) = (\{x_1, x_2, x_3, x_4, x_5\}, U)$, thus we see that $k = 2$ is the generalized decision optimal scale of S, and it is also the lower approximation optimal scale as well as the upper approximation optimal scale. However, the distribution optimal scale of S is $k = 1$. At the same time, we can check that the maximum distribution optimal scale of S is 1.

From Examples 2 and 3, we can see that there is no static relationship between the generalized decision optimal scale and the maximum distribution optimal scale.

Theorem 5. *Let* $S = (U, C \cup \{d\}) = \left(U, \{a_j^k \mid k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\} \cup \{d\}\right)$ *be an inconsistent multi-scale decision table which has I levels of granulations. For* $k \in \{1, 2, \ldots, I\}$, *then the following statements are equivalent:*

(1) $L_{C^k}(d) = L_{C^1}(d)$.
(2) $\mathrm{Bel}_{C^k}(d) = \mathrm{Bel}_{C^1}(d)$.
(3) $\sum_{j=1}^{r} \mathrm{Bel}_{C^k}(D_j) = \sum_{j=1}^{r} \mathrm{Bel}_{C^1}(D_j)$.

Proof

"(1)⇒(2)".

$$\begin{aligned} L_{C^k}(d) = L_{C^1}(d) &\Longrightarrow \underline{R_{C^k}}(D_j) = \underline{R_{C^1}}(D_j),\ \forall j \in \{1, 2, \ldots, r\}, \\ &\Longrightarrow P(\underline{R_{C^k}}(D_j)) = P(\underline{R_{C^1}}(D_j)),\ \forall j \in \{1, 2, \ldots, r\}, \\ &\Longrightarrow \mathrm{Bel}_{C^k}(D_j) = \mathrm{Bel}_{C^1}(D_j),\ \forall j \in \{1, 2, \ldots, r\}, \\ &\Longrightarrow \mathrm{Bel}_{C^k}(d) = \mathrm{Bel}_{C^1}(d). \end{aligned}$$

"(2)⇒(3)". It is obvious.

"(3)⇒(1)". Since $\sum_{j=1}^{r} \mathrm{Bel}_{C^k}(D_j) = \sum_{j=1}^{r} \mathrm{Bel}_{C^1}(D_j)$, we have

$$\sum_{j=1}^{r} |\underline{R_{C^k}}(D_j)| = \sum_{j=1}^{r} |\underline{R_{C^1}}(D_j)|. \tag{34}$$

Since

$$\underline{R_{C^k}}(D_j) \subseteq \underline{R_{C^1}}(D_j),\ \forall j \in \{1, 2, \ldots, r\}, \tag{35}$$

we have

$$|\underline{R_{C^k}}(D_j)| \leq |\underline{R_{C^1}}(D_j)|,\ \forall j \in \{1, 2, \ldots, r\}. \tag{36}$$

Hence, according to Eq. (36), Eq. (34) implies that

$$|\underline{R_{C^k}}(D_j)| = |\underline{R_{C^1}}(D_j)|,\ \forall j \in \{1, 2, \ldots, r\}. \tag{37}$$

In terms of Eq. (35), we must have

$$\underline{R_{C^k}}(D_j) = \underline{R_{C^1}}(D_j),\ \forall j \in \{1, 2, \ldots, r\}. \tag{38}$$

It follows that $L_{C^k}(d) = L_{C^1}(d)$. □

Theorem 5 shows that the kth level of scale is the lower approximation optimal scale of S if and only if it is the belief distribution optimal scale of S. Moreover, it can easily be conclude following

Theorem 6. *Let $S = (U, C \cup \{d\}) = \left(U, \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ be an inconsistent multi-scale decision table which has I levels of granulations. For $k \in \{1, 2, \ldots, I\}$, then the kth level of scale is the lower approximation optimal scale of S if and only if*

$$\sum_{j=1}^{r} \mathrm{Bel}_{C^k}(D_j) = \sum_{j=1}^{r} \mathrm{Bel}_{C^1}(D_j). \tag{39}$$

And (if there is $k + 1 \leq I$)

$$\sum_{j=1}^{r} \mathrm{Bel}_{C^{k+1}}(D_j) < \sum_{j=1}^{r} \mathrm{Bel}_{C^1}(D_j). \tag{40}$$

Theorem 7. *Let $S = (U, C \cup \{d\}) = \left(U, \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ be an inconsistent multi-scale decision table which has I levels of granulations. For $k \in \{1, 2, \ldots, I\}$, then the following statements are equivalent:*

(1) $H_{C^k}(d) = H_{C^1}(d)$.
(2) $\mathrm{Pl}_{C^k}(d) = \mathrm{Pl}_{C^1}(d)$.
(3) $\sum_{j=1}^{r} \mathrm{Pl}_{C^k}(D_j) = \sum_{j=1}^{r} \mathrm{Pl}_{C^1}(D_j)$.
(4) $\partial_{C^k}(x) = \partial_{C^1}(x),\ \forall x \in U$.

Proof
"(1)⇒(2)".

$$\begin{aligned} H_{C^k}(d) = H_{C^1}(d) &\Longrightarrow \overline{R_{C^k}}(D_j) = \overline{R_{C^1}}(D_j),\ \forall j \in \{1, 2, \ldots, r\}, \\ &\Longrightarrow P(\overline{R_{C^k}}(D_j)) = P(\overline{R_{C^1}}(D_j)),\ \forall j \in \{1, 2, \ldots, r\}, \\ &\Longrightarrow \mathrm{Pl}_{C^k}(D_j) = \mathrm{Pl}_{C^1}(D_j),\ \forall j \in \{1, 2, \ldots, r\}, \\ &\Longrightarrow \mathrm{Pl}_{C^k}(d) = \mathrm{Pl}_{C^1}(d). \end{aligned}$$

"(2)⇒(3)". It is obvious.

"(3)⇒(1)". Since $\sum_{j=1}^{r} \mathrm{Pl}_{C^k}(D_j) = \sum_{j=1}^{r} \mathrm{Pl}_{C^1}(D_j)$, we have

$$\sum_{j=1}^{r} |\overline{R_{C^k}}(D_j)| = \sum_{j=1}^{r} |\overline{R_{C^1}}(D_j)|. \tag{41}$$

By Eq. (23), we see that

$$\overline{R_{C^1}}(D_j) \subseteq \overline{R_{C^k}}(D_j), \ \forall j \in \{1, 2, \ldots, r\}, \tag{42}$$

then we have

$$\left|\overline{R_{C^1}}(D_j)\right| \leq \left|\overline{R_{C^k}}(D_j)\right|, \ \forall j \in \{1, 2, \ldots, r\}. \tag{43}$$

Hence, according to Eq. (43), Eq. (41) implies that

$$\left|\overline{R_{C^k}}(D_j)\right| = \left|\overline{R_{C^1}}(D_j)\right|, \ \forall j \in \{1, 2, \ldots, r\}. \tag{44}$$

In terms of Eq. (42), we must have

$$\overline{R_{C^k}}(D_j) = \overline{R_{C^1}}(D_j), \ \forall j \in \{1, 2, \ldots, r\}. \tag{45}$$

It follows that $H_{C^k}(d) = H_{C^1}(d)$.

"(4)⇒(1)". Assume that $\partial_{_{C^k}}(x) = \partial_{_{C^1}}(x)$ for all $x \in U$. For any $D_j \in U/R_d$ and $y \in U$, it is easy to see that $[y]_{C^k} \cap D_j \neq \emptyset$ if and only if $j \in \partial_{_{C^k}}(y)$, then we have

$$\begin{aligned} y \in \overline{R_{C^k}}(D_j) &\Longrightarrow [y]_{C^k} \cap D_j \neq \emptyset \\ &\Longrightarrow j \in \partial_{_{C^k}}(y) \\ &\Longrightarrow j \in \partial_{_{C^1}}(y) \\ &\Longrightarrow [y]_{C^1} \cap D_j \neq \emptyset \\ &\Longrightarrow y \in \overline{R_{C^1}}(D_j). \end{aligned}$$

Thus we have proved that

$$\overline{R_{C^k}}(D_j) \subseteq \overline{R_{C^1}}(D_j). \tag{46}$$

By Eq. (23), we then conclude that

$$\overline{R_{C^k}}(D_j) = \overline{R_{C^1}}(D_j).$$

"(1)⇒(4)". Assume that $H_{C^k}(d) = H_{C^1}(d)$, that is,

$$\overline{R_{C^k}}(D_j) = \overline{R_{C^1}}(D_j), \ \forall D_j \in U/R_d. \tag{47}$$

For any $x \in U$, since $R_{C^1} \subseteq R_{C^k}$, we have

$$\partial_{_{C^1}}(x) \subseteq \partial_{_{C^k}}(x). \tag{48}$$

On the on the hand, for $j \in V_d$, by Eq. (47), we have

$$\begin{aligned} j \in \partial_{_{C^k}}(x) &\Longrightarrow [x]_{C^k} \cap D_j \neq \emptyset \\ &\Longrightarrow x \in \overline{R_{C^k}}(D_j) \\ &\Longrightarrow x \in \overline{R_{C^1}}(D_j) \\ &\Longrightarrow [x]_{C^1} \cap D_j \neq \emptyset \\ &\Longrightarrow j \in \partial_{_{C^1}}(x). \end{aligned}$$

Hence

$$\partial_{_{C^k}}(x) \subseteq \partial_{_{C^1}}(x). \tag{49}$$

Combining Eqs. (48) and (49), we conclude $\partial_{_{C^k}}(x) = \partial_{_{C^1}}(x)$ for all $x \in U$. □

Theorem 7 shows that, in an inconsistent multi-scale decision table, the kth level of scale is the upper approximation optimal scale if and only if it is the plausibility distribution optimal scale if and only if it is the generalized decision optimal scale, in other words, all the upper approximation optimal scale, the plausibility distribution optimal scale, and the generalized decision optimal scale are the same. Similar to Theorem 6, we have following

Theorem 8. *Let $S = (U, C \cup \{d\}) = \left(U, \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ be an inconsistent multi-scale decision table which has I levels of granulations. For $k \in \{1, 2, \ldots, I\}$, then the kth level of scale is the upper approximation optimal scale of S if and only if*

$$\sum_{j=1}^{r} \mathrm{Pl}_{C^k}(D_j) = \sum_{j=1}^{r} \mathrm{Pl}_{C^1}(D_j). \tag{50}$$

And (if there is $k+1 \leq I$)

$$\sum_{j=1}^{r} \mathrm{Pl}_{C^{k+1}}(D_j) > \sum_{j=1}^{r} \mathrm{Pl}_{C^1}(D_j). \tag{51}$$

Theorem 9. *Let $S = (U, C \cup \{d\}) = \left(U, \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ be an inconsistent multi-scale decision table which has I levels of granulations. For $k \in \{1, 2, \ldots, I\}$, then*

$$H_{C^k}(d) = H_{C^1}(d) \Longrightarrow L_{C^k}(d) = L_{C^1}(d). \tag{52}$$

Proof. Assume that $H_{C^k}(d) = H_{C^1}(d)$, that is, $\overline{R_{C^k}}(D_j) = \overline{R_{C^1}}(D_j)$ for all $j \in \{1, 2, \ldots, r\}$. Then, for any $j \in \{1, 2, \ldots, r\}$ and $y \in U$, we have

$$[y]_{C^k} \cap D_j \neq \emptyset \Longleftrightarrow [y]_{C^1} \cap D_j \neq \emptyset. \tag{53}$$

For any $j \in \{1, 2, \ldots, r\}$ and $x \in U$, if $x \in \underline{R_{C^1}}(D_j)$, by the definition, we have

$$[x]_{C^1} \subseteq D_j. \tag{54}$$

Since $\{D_j | j = 1, 2, \ldots, r\}$ forms a partition of U, Eq. (54) means that

$$[y]_{C^1} \cap D_j \neq \emptyset \text{ and } [y]_{C^1} \cap D_i = \emptyset, \forall i \neq j. \tag{55}$$

Then, by Eq. (53), we conclude that

$$[y]_{C^k} \cap D_j \neq \emptyset \text{ and } [y]_{C^k} \cap D_i = \emptyset, \forall i \neq j. \tag{56}$$

Consequently, $[x]_{C^k} \subseteq D_j$, which follows that $x \in \underline{R_{C^k}}(D_j)$. Thus we have proved that $\underline{R_{C^1}}(D_j) \subseteq \underline{R_{C^k}}(D_j)$. On the other hand, by Eq. (23), we know that $\underline{R_{C^k}}(D_j) \subseteq \underline{R_{C^1}}(D_j)$. Therefore $\underline{R_{C^k}}(D_j) = \underline{R_{C^1}}(D_j)$ for all $j \in \{1, 2, \ldots, r\}$, i.e. $L_{C^k}(d) = L_{C^1}(d)$. □

Theorem 9 shows that if S^k is upper approximation consistent to S then it must be lower approximation consistent to S. Moreover, if k_1 is the upper approximation optimal scale of S and k_2 is the lower approximation optimal scale of S then $k_1 \leq k_2$, alternatively, the lower approximation optimal scale of S is, in general, not less than the upper approximation optimal scale of S. The next example shows that the converse of Theorem 9 is not always true.

Example 4. Table 7 gives an example of multi-scale decision table $S = (U, C \cup \{d\})$ which has 2 levels of granulations and one attribute a, where $U = \{x_1, x_2, \ldots, x_7\}$, $C = \{a\}$. It can be calculated that
$L_{a^1}(d) = L_{a^2}(d) = (\emptyset, \{x_2, x_4\}, \emptyset)$,
$H_{a^1}(d) = (\{x_1, x_3, x_5, x_6, x_7\}, \{x_2, x_3, x_4, x_5, x_6\}, \{x_1, x_7\})$,
$H_{a^2}(d) = (\{x_1, x_3, x_5, x_6, x_7\}, U, \{x_1, x_3, x_5, x_6, x_7\}) \neq H_{a^1}(d)$,
thus we see that $k = 2$ is the lower approximation optimal scale of S, but the upper approximation optimal scale of S is 1.

In summary, if we use k_l, k_u, k_d, k_m, k_b, k_p, and k_g to represent the lower approximation, the upper approximation, the distribution, the maximum distribution, the belief distribution, the plausibility distribution, and the generalized decision optimal scale of an inconsistent multi-scale decision table S, respectively, then according to Theorems 4-9, we have following equalities:

$$k_d \leq k_u = k_p = k_g \leq k_l = k_b, \text{ and } k_d \leq k_m.$$

Table 7
A multi-scale decision table.

U	a^1	a^2	d
x_1	1	S	1
x_2	3	L	2
x_3	2	S	2
x_4	3	L	2
x_5	2	S	1
x_6	2	S	1
x_7	1	S	3

4.3. Optimal scale selection in multi-scale decision tables with a probabilistic rough set model

In the Pawlak rough set model, the lower approximation is the union of those equivalence classes that are included in the set and the upper approximation is the union of those that have a non-empty overlap with the set. The rules induced by the lower approximation must be absolutely consistent or correct, namely, the classification must be completely correct or certain. However, the definitions of approximations do not allow any errors, which rarely happens in practice. So, several probabilistic rough set models were developed to solve these problems [5,21,20,52,59,64,65]. Probabilistic rough set approximations were formulated based on the notions of rough membership functions which can be interpreted in terms of conditional probabilities or a posteriori probabilities. Threshold values, known as parameters, are applied to a rough membership function to obtain probabilistic or parameterized approximations.

Let (U, R) be a Pawlak approximation space, $0 \leq \alpha < \beta \leq 1$, and $X \in \mathcal{P}(U)$, the standard rough approximations were extended to generalized probabilistic approximations by Yao and Wong [56]:

$$\begin{aligned} \underline{R}^{\beta}(X) &= \{x \in U | P(X|[x]_R) \geq \beta\}, \\ \overline{R}^{1-\alpha}(X) &= \{x \in U | P(X|[x]_R) > \alpha\}, \end{aligned} \tag{57}$$

The condition $\alpha < \beta$ ensures that the lower approximation is smaller than the upper approximation in order to be consistent with existing approximation operators. A pair of parameters (α, β) with $0 \leq \alpha < \beta \leq 1$ can be determined from a loss (cost) function within the decision-theoretic rough set model proposed by Yao and Wong [51,56].

With a pair of arbitrary α and β, the probabilistic approximation operators defined as Eq. (57) are not necessarily dual to each other. By setting $\alpha = 1 - \beta$, then, the lower and upper probabilistic approximation operators are dual operators. On the other hand, to ensure that the lower approximation is smaller than the upper approximation we set $\beta \in (0.5, 1]$, then

$$\begin{aligned} \underline{R}^{\beta}(X) &= \{x \in U | P(X|[x]_R) \geq \beta\}, \\ \overline{R}^{\beta}(X) &= \{x \in U | P(X|[x]_R) > 1 - \beta\}, \end{aligned} \tag{58}$$

$\underline{R}^{\beta}(X)$ and $\overline{R}^{\beta}(X)$ are called the β lower approximation and the β upper approximation of X w.r.t. (U, R) respectively, and the pair $(\underline{R}^{\beta}(X), \underline{R}^{\beta}(X))$ is called rough set with β-precision.

In the discussion to follow, we will use the dual probabilistic approximation operators.

Proposition 5. *Let (U, R) be a Pawlak approximation space and $\beta \in (0.5, 1]$, then $\underline{R_B}^{\beta}$ and $\overline{R_B}^{\beta}$ satisfy the following properties: for $X, Y \in \mathcal{P}(U)$,*

(1) $\underline{R}^{\beta}(X) = \sim \overline{R}^{\beta}(\sim X)$.
(2) $\underline{R}^{\beta}(X) \subseteq \overline{R}^{\beta}(X)$.
(3) $\underline{R}^{1}(X) = \underline{R}(X)$, $\overline{R}^{1}(X) = \overline{R}(X)$.
(4) $X \subseteq Y \Longrightarrow \underline{R}^{\beta}(X) \subseteq \underline{R}^{\beta}(Y)$, $\overline{R}^{\beta}(X) \subseteq \overline{R}^{\beta}(Y)$.
(5) $0.5 < \beta \leq \alpha \leq 1 \Longrightarrow \underline{R}^{\alpha}(X) \subseteq \underline{R}^{\beta}(X)$, $\overline{R}^{\beta}(X) \subseteq \overline{R}^{\alpha}(X)$.

Let $S = (U, C \cup \{d\}) = \left(U, \{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\} \cup \{d\}\right)$ be a multi-scale decision table which has I levels of granulations. For $1 \leq k \leq I$ and $\beta \in (0.5, 1]$, denote

$$\begin{aligned} L_{C^k}^{\beta}(d) &= \left(\underline{R_{C^k}}^{\beta}(D_1), \underline{R_{C^k}}^{\beta}(D_2), \ldots, \underline{R_{C^k}}^{\beta}(D_r)\right), \\ H_{C^k}^{\beta}(d) &= \left(\overline{R_{C^k}}^{\beta}(D_1), \overline{R_{C^k}}^{\beta}(D_2), \ldots, \overline{R_{C^k}}^{\beta}(D_r)\right), \\ \mathrm{Bel}_{C^k}^{\beta}(d) &= \sum_{j=1}^{r} \mathrm{Bel}_{C^k}^{\beta}(D_j) =: \sum_{j=1}^{r} \frac{|\underline{R_{C^k}}^{\beta}(D_j)|}{|U|}, \\ \mathrm{Pl}_{C^k}^{\beta}(d) &= \sum_{j=1}^{r} \mathrm{Pl}_{C^k}^{\beta}(D_j) =: \sum_{j=1}^{r} \frac{|\overline{R_{C^k}}^{\beta}(D_j)|}{|U|}. \end{aligned} \tag{59}$$

Definition 12. Let $S = (U, C \cup \{d\}) = \left(U, \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ be an inconsistent multi-scale decision table which has I levels of granulations. For $k \in \{1, 2, \ldots, I\}$ and $\beta \in (0.5, 1]$, we say that

(1) $S^k = (U, C^k \cup \{d\}) = \left(U, \left\{a_j^k | j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ is β lower approximation consistent to S if $L_{C^k}^{\beta}(d) = L_{C^1}^{\beta}(d)$. And, the kth level of scale is said to be the β lower approximation optimal scale if S^k is β lower approximation consistent to S and S^{k+1} (if there is $k + 1$) is not β lower approximation consistent to S.

(2) $S^k = (U, C^k \cup \{d\}) = \left(U, \left\{a_j^k | j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ is β upper approximation consistent to S if $H_{C^k}^{\beta}(d) = H_{C^1}^{\beta}(d)$. And, the kth level of scale is said to be the β upper approximation optimal scale if S^k is β upper approximation consistent to S and S^{k+1} (if there is $k+1$) is not β upper approximation consistent to S.

(3) $S^k = (U, C^k \cup \{d\}) = \left(U, \left\{a_j^k | j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ is β belief distribution consistent to S if $\mathrm{Bel}_{C^k}^{\beta}(d) = \mathrm{Bel}_{C^1}^{\beta}(d)$. And, the kth level of scale is said to be the β belief distribution optimal scale of S if S^k is β belief distribution consistent to S and S^{k+1} (if there is $k+1$) is not β belief distribution consistent to S.

(4) $S^k = (U, C^k \cup \{d\}) = \left(U, \left\{a_j^k | j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ is β plausibility distribution consistent to S if $\mathrm{Pl}_{C^k}^{\beta}(d) = \mathrm{Pl}_{C^1}^{\beta}(d)$. And, the kth level of scale is said to be the β plausibility distribution optimal scale of S if S^k is β plausibility distribution consistent to S and S^{k+1} (if there is $k+1$) is not β plausibility distribution consistent to S.

Theorem 10. *Let* $S = (U, C \cup \{d\}) = \left(U, \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ *be an inconsistent multi-scale decision table which has I levels of granulations. For* $k \in \{1, 2, \ldots, I\}$ *and* $\beta \in (0.5, 1]$, *then*

(1) $L_{C^k}^{\beta}(d) = L_{C^1}^{\beta}(d) \Longrightarrow \mathrm{Bel}_{C^k}^{\beta}(d) = \mathrm{Bel}_{C^1}^{\beta}(d)$.
(2) $H_{C^k}^{\beta}(d) = H_{C^1}^{\beta}(d) \Longrightarrow \mathrm{Pl}_{C^k}^{\beta}(d) = \mathrm{Pl}_{C^1}^{\beta}(d)$.

Proof. It follows directly from the definitions. □

Theorem 11. *Let* $S = (U, C \cup \{d\}) = \left(U, \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ *be an inconsistent multi-scale decision table which has I levels of granulations. For* $k \in \{1, 2, \ldots, I\}$ *and* $\beta \in (0.5, 1]$, *let*

$$\begin{aligned} \beta_{C^k} &= \min\left\{\max_{1 \le j \le r} P(D_j | [x]_{C^k}) | x \in U\right\}, \\ \beta_0 &= \min\{\beta_{C^k}, \beta_{C^1}\}. \end{aligned} \tag{60}$$

If $\beta_0 > 0.5$, *then*

(1) *For* $\beta \in (0.5, \beta_0]$, *if* S^k *is* β *lower approximation consistent to S, then* S^k *is maximum distribution consistent to S.*
(2) *For* $\beta \in (0.5, \beta_{C^1}]$, *if* S^k *is maximum distribution consistent to S, then* S^k *is* β *lower approximation consistent to S.*
(3) *For* $\beta \in (0.5, \beta_0]$, *the kth level of scale is the maximum distribution optimal scale of S if and only if k is the* β *lower approximation optimal scale of S.*

Proof

(1) If S^k is β lower approximation consistent to S and $\beta \in (0.5, \beta_0]$, then $\underline{R_{C^k}}^{\beta}(D_j) = \underline{R_{C^1}}^{\beta}(D_j)$ for all $j \in \{1, 2, \ldots, r\}$. For any $x \in U$, we know from $\beta_0 > 0.5$ that both $\gamma_{C^1}(x)$ and $\gamma_{C^k}(x)$ are singletons. Then, we have

$$\begin{aligned} D_j \in \gamma_{C^1}(x) &\Longrightarrow P(D_j | [x]_{C^1}) \ge \beta_0 \Longrightarrow P(D_j | [x]_{C^1}) \ge \beta \\ &\Longrightarrow x \in \underline{R_{C^1}}^{\beta}(D_j) \qquad \Longrightarrow x \in \underline{R_{C^k}}^{\beta}(D_j) \\ &\Longrightarrow P(D_j | [x]_{C^k}) \ge \beta \quad \Longrightarrow D_j \in \gamma_{C^k}(x), \end{aligned}$$

and

$$\begin{aligned} D_j \in \gamma_{C^k}(x) &\Longrightarrow P(D_j | [x]_{C^k}) \ge \beta_0 \Longrightarrow P(D_j | [x]_{C^k}) \ge \beta \\ &\Longrightarrow x \in \underline{R_{C^k}}^{\beta}(D_j) \qquad \Longrightarrow x \in \underline{R_{C^1}}^{\beta}(D_j) \\ &\Longrightarrow P(D_j | [x]_{C^1}) \ge \beta \quad \Longrightarrow D_j \in \gamma_{C^1}(x). \end{aligned}$$

Therefore $\gamma_{C^k}(x) = \gamma_{C^1}(x)$ for all $x \in U$, which follows that S^k is maximum distribution consistent to S.

(2) Assume that S^k is maximum distribution consistent to S, then for any $x \in U$, we have $\gamma_{C^k}(x) = \gamma_{C^1}(x)$. Denote

$$\mathcal{J}([x]_{C^k}) = \{[y]_{C^1} \in U/R_{C^1} | [y]_{C^1} \subseteq [x]_{C^k}\}.$$

Since $R_{C^1} \subseteq R_{C^k}$, it can easily be verified that $\mathcal{J}([x]_{C^k})$ forms a partition of $[x]_{C^k}$.
For any $j \in \{1, 2, \ldots, r\}$, we have

$$\begin{aligned} x \in \underline{R_{C^1}}^{\beta}(D_j) &\Longrightarrow P(D_j|[x]_{C^1}) \geq \beta > 0.5 \\ &\Longrightarrow \gamma_{c^1}(x) = \{D_j\} \\ &\Longrightarrow \gamma_{c^k}(x) = \{D_j\}. \end{aligned}$$

If $[y]_{C^1} \in \mathcal{J}([x]_{C^k})$, then $\gamma_{c^k}(x) = \gamma_{c^1}(y)$. By the assumption that $\gamma_{c^1}(y) = \gamma_{c^k}(y)$, we then have $\gamma_{c^1}(y) = \gamma_{c^k}(x) = \{D_j\}$, which implies that $P(D_j|[y]_{C^1}) \geq \beta_{c^1} > \beta$. Thus we have

$$\begin{aligned} P(D_j|[x]_{C^k}) &= \left(\sum\{|[y]_{C^1} \cap D_j| | [y]_A \in \mathcal{J}([x]_{C^k})\}\right) / |[x]_C^k| \\ &= \sum\left\{P(D_j|[y]_{C^1}) \cdot \frac{|[y]_{C^1}|}{|[x]_{C^k}|} | [y]_{C^1} \in \mathcal{J}([x]_{C^k})\right\} \\ &\geq \beta \cdot \sum\left\{\frac{|[y]_{C^1}|}{|[x]_{C^k}|} | [y]_{C^1} \in \mathcal{J}([x]_{C^k})\right\} = \beta. \end{aligned}$$

Consequently $x \in \underline{R_{C^k}}^{\beta}(D_j)$, from which follows that

$$\underline{R_{C^1}}^{\beta}(D_j) \subseteq \underline{R_{C^k}}^{\beta}(D_j). \tag{61}$$

Conversely, for any $x \in U$, since

$$\begin{aligned} x \in \underline{R_{C^k}}^{\beta}(D_j) &\Longrightarrow P(D_j|[x]_{C^k}) \geq \beta > 0.5 \\ &\Longrightarrow \gamma_{c^k}(x) = \{D_j\} \\ &\Longrightarrow \gamma_{c^1}(x) = \{D_j\}, \end{aligned}$$

we have

$$P(D_j|[x]_{C^1}) = \max_{i \in \{1,2,\ldots,r\}} P(D_i|[x]_{C^1}) \geq \beta_{c^1} > \beta. \tag{62}$$

Then, by the definition, we have $x \in \underline{R_{C^1}}^{\beta}(D_j)$, and consequently,

$$\underline{R_{C^k}}^{\beta}(D_j) \subseteq \underline{R_{C^1}}^{\beta}(D_j). \tag{63}$$

Thus, from Eqs. (61) and (63), we have proved that $\underline{R_{C^k}}^{\beta}(D_j) = \underline{R_{C^1}}^{\beta}(D_j)$ for all $j \in \{1, 2, \ldots, r\}$, that is, S^k is β lower approximation consistent to S.

(3) For $\beta \in (0.5, \beta_0]$, we can conclude from (1) and (2) that S^k is maximum distribution consistent to S if and only if S^k is β lower approximation consistent to S. Therefore k is the maximum distribution optimal scale of S if and only if k is the β lower approximation optimal scale of S. □

Theorem 12. *Let $S = (U, C \cup \{d\}) = \left(U, \left\{a_j^k | k = 1, 2, \ldots, I, j = 1, 2, \ldots, m\right\} \cup \{d\}\right)$ be an inconsistent multi-scale decision table which has I levels of granulations. For $k \in \{1, 2, \ldots, I\}$ and $\beta \in (0.5, 1]$, let*

$$\begin{aligned} \alpha_{c^k} &= \min\left\{P(D_j|[x]_{C^k}) | x \in U, D_j \cap [x]_{C^k} \neq \emptyset\right\} \\ &= \min\{P(D_j|[x]_{C^k}) | x \in U, P(D_j|[x]_{C^k}) > 0\}, \\ \alpha_0 &= \min\{\alpha_{c^1}, \alpha_{c^k}\}. \end{aligned} \tag{64}$$

Then

(1) *For $\beta \in (1 - \alpha_0, 1]$, if S^k is generalized decision consistent to S, then S^k is β upper approximation consistent to S.*
(2) *For $\beta \in (1 - \alpha_{c^k}, 1]$, if S^k is β upper approximation consistent to S, then S^k is generalized decision consistent to S.*
(3) *For $\beta \in (1 - \alpha_0, 1]$, k is the generalized decision optimal scale of S if and only if k is the β upper approximation optimal scale of S.*

Proof

(1) If S^k is generalized decision consistent to S, that is, $\partial_{c^k}(x) = \partial_{c^1}(x)$ for all $x \in U$. Then, for any $j \in \{1, 2, \ldots, r\}$, notice that

$$j \in \partial_{c^k}(x) \Longleftrightarrow D_j \cap [x]_{C^k} \neq \emptyset \Longleftrightarrow P(D_j|[x]_{C^k}) > 0, \tag{65}$$

we have

$$\begin{aligned} x \in \overline{R_{C^k}}^{\beta}(D_j) &\Longrightarrow P(D_j|[x]_{C^k}) > 1-\beta \Longrightarrow D_j \cap [x]_{C^k} \neq \emptyset \\ &\Longrightarrow j \in \partial_{C^k}(x) \Longrightarrow j \in \partial_{C^1}(x) \\ &\Longrightarrow D_j \cap [x]_{C^1} \neq \emptyset \Longrightarrow P(D_j|[x]_{C^1}) \geq \alpha_{C^1} \geq \alpha_0 > 1-\beta \\ &\Longrightarrow x \in \overline{R_{C^1}}^{\beta}(D_j). \end{aligned}$$

Conversely,

$$\begin{aligned} x \in \overline{R_{C^1}}^{\beta}(D_j) &\Longrightarrow P(D_j|[x]_{C^1}) > 1-\beta \Longrightarrow j \in \partial_{C^1}(x) \\ &\Longrightarrow j \in \partial_{C^k}(x) \Longrightarrow D_j \cap [x]_{C^k} \neq \emptyset \\ &\Longrightarrow P(D_j|[x]_{C^k}) \geq \alpha_{C^k} \geq \alpha_0 > 1-\beta \\ &\Longrightarrow x \in \overline{R_{C^k}}^{\beta}(D_j). \end{aligned}$$

Thus we have proved that $\overline{R_{C^k}}^{\beta}(D_j) = \overline{R_{C^1}}^{\beta}(D_j)$ for all $j \in \{1, 2, \ldots, r\}$, that is, S^k is β upper approximation consistent to S.

(2) Assume that S^k is β upper approximation consistent to S. For $\beta \in (1-\alpha_{C^k}, 1]$, by the definition, we have $\overline{R_{C^k}}^{\beta}(D_j) = \overline{R_{C^1}}^{\beta}(D_j)$ for all $j \in \{1, 2, \ldots, r\}$. Then, for any $x \in U$,

$$\begin{aligned} j \in \partial_{C^k}(x) &\Longrightarrow P(D_j|[x]_{C^k}) \geq \alpha_{C^k} > 1-\beta \\ &\Longrightarrow x \in \overline{R_{C^k}}^{\beta}(D_j) \Longrightarrow x \in \overline{R_{C^1}}^{\beta}(D_j) \\ &\Longrightarrow P(D_j|[x]_{C^1}) > 1-\beta \Longrightarrow j \in \partial_{C^1}(x). \end{aligned}$$

It follows that $\partial_{C^k}(x) \subseteq \partial_{C^1}(x)$. On the other hand, by Eq. (33) we see that $\partial_{C^1}(x) \subseteq \partial_{C^k}(x)$. Thus we have proved that $\partial_{C^1}(x) = \partial_{C^k}(x)$ for all $x \in U$, that is, S^k is generalized decision consistent to S.

(3) For $\beta \in (1-\alpha_0, 1]$, we can see from (1) and (2) that S^k is generalized decision consistent to S if and only if S^k is β upper approximation consistent to S. Therefore k is the generalized decision optimal scale of S if and only if k is the β upper approximation optimal scale of S. □

5. Conclusions

In rough-set data analysis, each object can only take on one value under each attribute in almost all information tables. However, in some real-life applications, one has to make decision with data measured at different levels of granulations. That is, an object may take on different values under the same attribute, depending on at which scale it is measured. In this paper, we have introduced the concept of multi-scale information table from the perspective of granular computation which has different levels of granulations. In such a system, each object under each attribute is represented by different scales at different levels of granulations having a granular information transformation from a finer to a coarser labelled value. We have also defined lower and upper approximations with reference to different levels of granulations in multi-scale information tables and examined their properties. We have further discussed optimal scale selection with various requirements in multi-scale decision tables with the standard rough set model and a dual probabilistic rough set model, where the rough membership functions and the belief functions are employed to measure uncertainty. With reference to the optimal levels of granulations, one can analyze corresponding knowledge acquisition in the sense of rule induction in multi-scale decision tables. Finally, we have presented relationship among different notions of optimal scales in multi-scale information tables. For further study, new approaches to granular representation and new models for knowledge acquisition in complicated multi-scale information tables such as incomplete information tables, fuzzy information tables, set-valued information tables, interval-valued information tables need to be formulated in future studies.

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (Nos. 61272021, 61075120, 11071284, 61202206, and 61173181), the Zhejiang Provincial Natural Science Foundation of China (No. LZ12F03002) and the Geographical Modeling and Geocomputation Program under the Focused Investment Scheme of The Chinese University of Hong Kong.

References

[1] A. Bargiela, W. Pedrycz, Granular Computing: An Introduction, Kluwer Academic Publishers, Boston, 2002.
[2] A. Bargiela, W. Pedrycz, Toward a theory of granular computing for human-centered information processing, IEEE Transactions on Fuzzy Systems 16 (2008) 320–330.
[3] T. Bittner, B. Smith, A theory of granular partitions, in: M. Duckham, M.F. Goodchild, M.F. Worboys (Eds.), Foundations of Geographic Information Science, Taylor & Francis, London, 2003, pp. 117–151.
[4] S. Greco, B. Matarazzo, R. Slowinski, Rough approximation by dominance relation, International Journal of Intelligent Systems 17 (2002) 153–171.
[5] J.W. Grzymala-Busse, Y.Y. Yao, Probabilistic rule induction with the LERS data mining system, International Journal of Intelligent Systems 26 (2011) 518–539.
[6] J.W. Grzymala-Busse, X. Zuo, Classification strategies using certain and possible rules, in: L. Polkowski, A. Skowron (Eds.), Rough Sets and Current Trends in Computing, Lecture Notes in Computer Science, vol. 1424, 1998, pp. 37–44.
[7] A.E. Hassanien, Rough set approach for attribute reduction and rule generation: A case of patients with suspected breast cancer, Journal of the American Society for Information Science and Technology 55 (2004) 954–962.
[8] T.P. Hong, L.-H. Tseng, S.-L. Wang, Learning rules from incomplete training examples by rough sets, Expert Systems with Applications 22 (2002) 285–293.
[9] Q.H. Hu, J.F. Liu, D.R. Yu, Mixed feature selection based on granulation and approximation, Knowledge-Based Systems 21 (2008) 294–304.
[10] M. Inuiguchi, S. Hirano, S. Tsumoto (Eds.), Rough Set Theory and Granular Computing, Springer, Berlin, 2002
[11] C.M. Keet, A Formal Theory of Granularity, Ph.D. Thesis, KRDB Research Centre, Faculty of Computer Science, Free University of Bozen-Bolzano, Italy, 2008.
[12] J. Komorowski, Z. Pawlak, L. Polkowski, A. Skowron, Rough sets: tutorial, in: S.K. Pal, A. Skowron (Eds.), Rough Fuzzy Hybridization, A New Trend in Decision Making, Springer, Berlin, 1999, pp. 3–98.
[13] Y. Leung, D.Y. Li, Maximal consistent block technique for rule acquisition in incomplete information systems, Information Sciences 153 (2003) 85–106.
[14] Y. Leung, W.-Z. Wu, W.-X. Zhang, Knowledge acquisition in incomplete information systems: A rough set approach, European Journal of Operational Research 168 (2006) 164–180.
[15] Y. Leung, J.S. Zhang, Z.B. Xu, Clustering by scale-space filtering, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 1396–1410.
[16] D.Y. Li, B. Zhang, Y. Leung, On knowledge reduction in inconsistent decision information systems, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 12 (2004) 651–672.
[17] T.Y. Lin, Granular computing: from rough sets and neighborhood systems to information granulation and computing with words, in: European Congress on Intelligent Techniques and Soft Computing, September 8–12, 1997, pp. 1602–1606.
[18] T.Y. Lin, Y.Y. Yao, L.A. Zadeh (Eds.), Data Mining, Rough Sets and Granular Computing, Physica-Verlag, Heidelberg, 2002
[19] P.J. Lingras, Y.Y. Yao, Data mining using extensions of the rough set model, Journal of the American Society for Information Science 49 (1998) 415–422.
[20] D. Liu, T.R. Li, D. Ruan, Probabilistic model criteria with decision-theoretic rough sets, Information Sciences 181 (2011) 3709–3722.
[21] W.M. Ma, B.Z. Sun, Probabilistic rough set over two universes and rough entropy, International Journal of Approximate Reasoning 53 (2012) 608–619.
[22] J.-M. Ma, W.-X. Zhang, Y. Leung, X.-X. Song, Granular computing and dual Galois connection, Information Sciences 177 (2007) 5365–5377.
[23] J.-S. Mi, W.-Z. Wu, W.-X. Zhang, Approaches to knowledge reduction based on variable precision rough sets model, Information Sciences 159 (2004) 255–272.
[24] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers., Boston, 1991.
[25] Z. Pawlak, Drawing conclusions from data—the rough set way, International Journal of Intelligent Systems 16 (2001) 3–11.
[26] Z. Pawlak, A. Skowron, Rough membership functions, in: R.R. Yager, M. Fedrizzi, J. Kacprzyk (Eds.), Advances in the Dempster–Shafer Theory of Evidence, John Wiley and Sons, New York, 1994, pp. 251–271.
[27] Z. Pawlak, S.K.M. Wong, W. Ziarko, Rough sets: probabilistic versus deterministic approach, International Journal of Man–Machine Studies 29 (1988) 81–95.
[28] W. Pedrycz (Ed.), Granular Computing: An Emerging Paradigm, Physica-Verlag, Heidelberg, 2001
[29] W. Pedrycz, A. Skowron, V. Kreinovich (Eds.), Handbook of Granular Computing, Wiley, 2008
[30] L. Polkowski, S. Tsumoto, T.Y. Lin (Eds.), Rough Set Methods and Applications, Physica-Verlag, Berlin, 2000
[31] Y.H. Qian, J.Y. Liang, C.Y. Dang, Incomplete multigranulation rough set, IEEE Trasactions on Systems, Man and Cybernetics—Part A 40 (2010) 420–431.
[32] Y.H. Qian, J.Y. Liang, Y.Y. Yao, C.Y. Dang, MGRS: A multi-granulation rough set, Information Sciences 180 (2010) 949–970.
[33] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, Princeton, 1976.
[34] M.-W. Shao, W.-X. Zhang, Dominance relation and rules in an incomplete ordered information system, International Journal of Intelligent Systems 20 (2005) 13–27.
[35] A. Skowron, The relationship between rough set theory and evidence theory, Bulletin of the Polish Academy of Sciences: Mathematics 37 (1989) 87–90.
[36] A. Skowron, The rough sets theory and evidence theory, Fundamenta Informaticae 13 (1990) 245–262.
[37] A. Skowron, J. Grzymala-Busse, From rough set theory to evidence theory, in: R.R. Yager, M. Fedrizzi, J. Kacprzyk (Eds.), Advance in the Dempster–Shafer Theory of Evidence, Wiley, New York, 1994, pp. 193–236.
[38] A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, in: R. Slowinski (Ed.), Intelligent Decision Support—Handbook of Applications and Advances of the Rough Sets Theory, Kluwer Academic Publishers., Boston, 1992, pp. 331–362.
[39] A. Skowron, J. Stepaniuk, Information granules: Towards foundations of granular computing, International Journal of Intelligent Systems 16 (2001) 57–85.
[40] A. Skowron, J. Stepaniuk, R.W. Swiniarski, Modeling rough granular computing based on approximation spaces, Information Sciences 184 (2012) 20–43.
[41] W.-Z. Wu, Attribute reduction based on evidence theory in incomplete decision systems, Information Sciences 178 (2008) 1355–1371.
[42] W.-Z. Wu, Y. Leung, Theory and applications of granular labelled partitions in multi-scale decision tables, Information Sciences 181 (2011) 3878–3897.
[43] W.-Z. Wu, Y. Leung, J.-S. Mi, Granular computing and knowledge reduction in formal contexts, IEEE Transactions on Knowledge and Data Engineering 21 (2009) 1461–1474.
[44] W.-Z. Wu, Y. Leung, W.-X. Zhang, Connections between rough set theory and Dempster–Shafer theory of evidence, International Journal of General Systems 31 (2002) 405–430.
[45] W.-Z. Wu, M. Zhang, H.-Z. Li, J.-S. Mi, Knowledge reduction in random information systems via Dempster–Shafer theory of evidence, Information Sciences 174 (2005) 143–164.
[46] R.R. Yager, Intelligent social network analysis using granular computing, International Journal of Intelligent Systems 23 (2008) 1196–1219.
[47] J.T. Yao, Recent developments in granular computing: a bibliometrics study, in: Proceedings of IEEE International Conference on Granular Computing, Hangzhou, China, Aug 26–28, 2008, pp. 74–79.
[48] Y.Y. Yao, Stratified rough sets and granular computing, in: R.N. Dave, T. Sudkamp (Eds.), Proceedings of the 18th International Conference of the North American Fuzzy Information Processing Society, IEEE Press, 1999, pp. 800–804.
[49] Y.Y. Yao, Information granulation and rough set approximation, International Journal of Intelligent Systems 16 (2001) 87–104.
[50] Y.Y. Yao, Information granulation and approximation in a decision-theoretic model of rough sets, in: L. Polkowski, S.K. Pal, A. Skowron (Eds.), Rough-neuro Computing: A Way to Computing with Words, Springer, Berlin, 2003, pp. 491–518.
[51] Y.Y. Yao, Three-way decisions with probabilistic rough sets, Information Sciences 180 (2010) 341–353.
[52] Y.Y. Yao, Probabilistic rough set approximations, International Journal of Approximate Reasoning 49 (2008) 255–271.
[53] Y.Y. Yao, A partition model of granular computing, Transactions on Rough Sets I, Lecture Notes in Computer Science 3100 (2004) 232–253.
[54] Y.Y. Yao, C.-J. Liau, N. Zhong, Granular computing based on rough sets, quotient space theory, and belief functions, Lecture Notes in Computer Science 2871 (2003) 152–159.
[55] Y.Y. Yao, P.J. Lingras, Interpretations of belief functions in the theory of rough sets, Information Sciences 104 (1998) 81–106.
[56] Y.Y. Yao, S.K.M. Wong, A decision theoretic framework for approximating concepts, International Journal of Man-Machine Studies 37 (1992) 793–809.
[57] L.A. Zadeh, Fuzzy sets and information granularity, in: N. Gupta, R. Ragade, R.R. Yager (Eds.), Advances in Fuzzy Set Theory and Applications, North-Holland, Amsterdam, 1979, pp. 3–18.

[58] L.A. Zadeh, Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets and Systems 90 (1997) 111–127.
[59] H.Y. Zhang, J. Zhou, D.Q. Miao, C. Gao, Bayesian rough set model: A further investigation, International Journal of Approximate Reasoning 53 (2012) 541–557.
[60] W.-X. Zhang, Y. Leung, W.-Z. Wu, Information Systems and Knowledge Discovery (in Chinese), Science Press, Beijing, 2003.
[61] W.-X. Zhang, J.-S. Mi, W.-Z. Wu, Approaches to knowledge reductions in inconsistent systems, International Journal of Intelligent Systems 18 (2003) 989–1000.
[62] Y.-Q. Zhang, Constructive granular systems with universal approximation and fast knowledge discovery, IEEE Transactions on Fuzzy Systems 13 (2005) 48–57.
[63] Y. Zhao, Y.Y. Yao, F. Luo, Data analysis based on discernibility and indiscernibility, Information Sciences 177 (2007) 4959–4976.
[64] W. Ziarko, Variable precision rough set model, Journal of Computer and System Science 46 (1993) 39–59.
[65] W. Ziarko, Probabilistic approach to rough sets, International Journal of Approximate Reasoning 49 (2008) 272–284.