KAPITEL 7 / *CHAPTER 7* [7]
THE METHODOLOGY OF SYNTHESIS OF INFORMATION TECHNOLOGY FOR STRUCTURING OF ROUGH DATA AND EXPERT KNOWLEDGE

DOI: 10.30890/2709-2313.2023-25-00-001

Introduction

It is very common when data or knowledge is not precise. Imprecision reflects the essence (content) of the statement (information, knowledge) and depends on the detail of the language used to describe it. For example, a statement like "the event happened two or three hours ago" is imprecise; a statement of the type: "the event occurred two hours ago" is precise if we are talking about a period of time and not accurate if it is necessary to indicate the exact time of the event.

Imprecision occurs in situations where the values of some parameters are measured with a predetermined error, it is this error that gives rise to inaccuracy.

The imprecision of the value means that the value can be obtained with an precise that does not exceed a certain threshold determined by the nature of the corresponding parameter of the object [15]. As an example of an imprecise value, it can be cited the weight of a product unit (for example, 600g ± 3 %) or the obtained measurement data of some characteristic of the object (for example, t = 2.5 ± 0.1s means that the true value of the quantity t lies in the interval from 2.4s to 2.6s).

In [15], the author notes that each imprecise value is characterized by a certain granularity ("graininess"), which reflects the degree of imprecision of the studied parameter of the object in relation to the size of the granule (grain). By its nature, the granule is a whole (indivisible) structure and can be used as a unit of measurement of the investigated imprecise value. In [15] attention is focused on the fact that an imprecise value is expressed as an integer, for example, the weight of an object of 1000 ± 1kg can be represented as 1 ton or 10 centners.

The degree of granularity varies from fine-grained to coarse-grained. Reducing the size of the granule (narrowing the context) naturally leads to an increase in the

[7]**Authors:** *Shved Alyona Volodymyrivna*

accuracy of the value of the analyzed parameter. In real conditions, the minimum limit degree of granularity is usually unattainable, respectively, as well as absolute accuracy. For example, consider the "blue" granule, when reducing the degree of granularity, we will get different shades of blue (which are 180 tones in the Pantone palette), for example, "light blue", "cold blue", etc. By decreasing the degree of granularity, we will get less and less noticeable values, at the same time, when increasing the degree of granularity, several less noticeable values (objects) are replaced by a larger (inaccurate) value, for example, the colors "indigo" and gray-blue will be identified with the granule " blue", although in fact they are different colors.

To model inaccuracy (in the above sense), the techniques to interval data, fuzzy set theory methods, and possibility theory methods have become widely used.

To present and process imprecise knowledge in information systems, assuming that knowledge is reflected in the classification of relevant elements (objects of the real or abstract world), methods of rough set theory have been widely used [16, 17, 23].

For example, let the initial set of elements of knowledge X is represented by two classes: $X_{S1} = \{X_1, X_2, X_3, X_4\}$ and $X_{S2} = \{X_5, X_6, X_7\}$, $X_{S1}, X_{S2} \subset X$, and let some target subset of knowledge $X_0 = \{X_3, X_4, X_7, X_8\}$, which must be assigned to one of the specified classes, is obtained. However, it is possible to see that elements $(X_3, X_4) \in S_1$, $X_7 \in S_2$, and the element $X_8 \notin (S_1, S_2)$. This characterizes the existence of a kind of "imprecise" classification. Rough set theory was proposed to analyze such situations [17, 18].

7.1. The basic aspects of rough set theory

Rough set theory (RST) offers a mathematical apparatus capable of correctly processing large arrays of unordered data and, based on the results of such processing, obtaining new knowledge [16, 17, 23]. This theory is based on the fact that knowledge is deeply embedded in people's ability to classify subjects, phenomena, objects, situations, etc.

Therefore, knowledge in the theory of rough sets is necessarily associated with a set of classification samples related to specific parts of the real or abstract world, which is called the universe of discourse (or, in short, the universe) [16, 17, 23].

Let $U \neq \varnothing$ be a finite set (universe) of the analyzed objects. Any subset of the universe $X \subseteq U$ is called a category on *U*, and any family of categories over *U* is knowledge. RST is based on the categories that form the classification of the given universe *U*, that is, on such a family $C = \{X_1, X_2, ..., X_n\}$, that $X_i \subseteq U$, $X_i \neq \varnothing$ for $i \neq j$, ($i, j = \overline{1,n}$) and $\bigcup X_i = U$. Such a family was named the knowledge base on *U*, which represents a set of basic aspects of classification (color, temperature, etc.) [16, 17, 23].

RST proposes to perform classification procedures based on equivalence relations, which are simpler than decision rules [16, 17, 23].

If *R* is an equivalence relation on *U*, then *IND*(*R*) denotes the family of equivalence classes (categories) of elements of *U*, and $[x]_R$ – denotes the category in *R* containing the element $x \in U$.

Then the knowledge base (KB) is the relational system $K = (U, R)$, where $U \neq \varnothing$ is a finite set of elements, *R* is the family of equivalence relations on *U*.

If we consider the target set of elements $X \in U$, the following situations can be considered in relation to the *IND*(*R*) classification [16, 17, 23]:

1. The set *X* is the union of some *R*-basic categories. In this case, the set *X* is exact set (*R*-definable).

2. The set *X* cannot be expressed as a union of some *R*-basic categories. In this case, the set *X* is *R*-rough or *R*-undefinable.

3. The lower approximation of *X* is the subset of *U* that can be classified with certainty as belonging to the target set *X*:

$$\underline{R}X = \{x \in U : [x]_R \subseteq X\}, \text{ or } x \in \underline{R}X, \text{ if and only if } [x]_R \subseteq X. \quad (1)$$

The lower approximation of the target set *X* is called the *R*-positive region of *X*:

$$POS_R(X) = \underline{R}X. \quad (2)$$

4. The upper approximation of X is the subset of U that can possibly be classified as belonging to the target set X:

$$\overline{R}X = \{x \in U : [x]_R \cap X \neq \varnothing\}, \text{ or } x \in \overline{R}X, \text{ if and only if } [x]_R \cap X \neq \varnothing. \quad (3)$$

5. The negative region of X is the subset of U that definitely do not belong to X:

$$NEG_R(X) = U - \overline{R}X. \quad (4)$$

6. The boundary region is the of U that belong to the upper approximation, but do not belong to the lower approximation of the target set X:

$$BN_R(X) = \overline{R}X - \underline{R}X. \quad (5)$$

RST allows modeling the uncertainty regarding the belonging of some elements of the universe to a given target set and makes it possible to estimate the degree of this uncertainty by introducing specific lower and upper approximations of this set.

7.2. Statement the problem of presentation and structuring of imprecise (rough) data and expert knowledge

Let $U \neq \varnothing$ be a finite set of analyzed objects (universe of elements). Based on U set, it is possible to distinguish subsets of elements of the universe $X_s \subseteq U$ (a concept or category in U), then any family of concepts in U is considered abstract knowledge about U. Thus, concepts form a partition (classification) of a given universe U, that is, in U it is possible select such a family $C = \{X_s \mid s = \overline{1,n}\}$ that, $X_s \subseteq U$, $X_s \neq \varnothing$, $X_s \cap X_t = \varnothing$ for $s \neq t$, $s,t = \overline{1,n}$, $\cup U_s = U$. A family of classifications on U forms a KB on U. Such a KB represents a set of aspects of the classification of objects of the universe.

Then the existing knowledge system can be presented in the form of KB $K = (U, R)$, where $U \neq \varnothing$ is a finite set of analyzed objects (universe of elements), R is an equivalence relation, on the basis of which equivalence classes ($IND(R)$) of elements of U can be formed. Each equivalence class (category) contains elements that have the

same properties (attributes). Within each such category, the elements are considered indistinguishable. Then by U/R we denote the family of all equivalence classes (classifications of U), and by $[u]_R$ denote the category in R containing the element $u \in U$.

A relational system (table) is used to graphically represent the knowledge system K. The rows of such table correspond to the analyzed objects – the elements $u \in U$, and the columns correspond to the features (criteria, attributes) of these elements. In the cell at the intersection of the j-th row and the l-th column, the value of the l-th attribute for the j-th element is displayed, thus each row of the table displays one element (object) of the universe and its corresponding attribute values. Such a table within the RST notation was called an information system (IS).

The concept of "information system" was introduced in [16, 17].

Let be given an IS $S = (U, A, V, f)$, where $U = \{u_j \mid j = \overline{1,z}\}$ is a non-empty finite set of elements (universe); $A = \{a_l \mid l = \overline{1,q}\}$ is a non-empty finite set of primitive attributes; $V = \bigcup_{a_l \in A} V_{a_l}$, V_{a_l} is the set of values of the attribute a_l (the area of the attribute a_l); $f \wedge U \times A \to V$ is an information function such that $\forall a_l \in A, u \in U, f(u, a_l) \in V_{a_l}$. Each subset $B \subseteq A$ is called an attribute. Attribute B can be both primitive ($|B| = 1$) and composite.

There is a one-to-one relationship between the concepts of KB and IS. Therefore, an arbitrary KB $K = (U, R)$ can be matched to the IS $S = (U, A, V, f)$, where each equivalence relation in the KB is represented in the IS by an attribute, and each equivalence class is represented by the value (values) of the attribute. All operations performed in KB based on equivalence relations can be performed in IS based on the separation of elements by sets of relevant attributes.

7.3. Synthesis of information technology for structuring of imprecise (rough) data and expert knowledge

The proposed information technology (IT) is designed to solve the problem of analysis and structuring of imprecise (rough, unordered) data and knowledge, and synthesis of new knowledge. The proposed IT can also be used when solving the problem of classification of raw data sets under imprecision, inconsistency, incompleteness of the source information (data, knowledge).

Let's consider the main ideas of IT for analysis and structuring of data and knowledge of IS, formed under imprecision, inconsistency, incompleteness of source information.

Let's assume that the *KB* $K = (U, R)$ is given, where U is the universe of elements, R is the equivalence relation. Let us correspond to the given KB the IS $S = (U, A, V, f)$, where $U = \{u_j \mid j = \overline{1,z}\}$ is a non-empty finite set of elements (universe); $A = \{a_l \mid l = \overline{1,q}\}$ is a non-empty finite set of primitive attributes; $V = \bigcup_{a_l \in A} V_{a_l}$, V_{a_l} is the set of values of the al attribute (the area of the al attribute); $f \wedge U \times A \to V$ is an information function.

The generalized structure of IT for structuring (classification) of imprecise (rough) data and knowledge of IS is shown in fig. 1.

The methodology of synthesis of IT for structuring (classification) of imprecise (rough) data and knowledge of IS using RST methods can be formally presented in the form of the following successive stages:

1. Problem statement, determination of initial conditions and limitations.

At this stage, a universe $U = \{u_j \mid j = \overline{1,z}\}$ is specified, the elements of which are characterized by a given set of features $A = \{a_l \mid l = \overline{1,q}\}$, sets of target subsets, $U_s \subseteq U$, $s < z$, are formed. The way of obtaining and the form of representation of the values of attributes $A = \{a_l \mid l = \overline{1,q}\}$ are determined.

The task is to determine the elements of the universe u_j that can be classified as

belonging to a given target set $U_s \subseteq U$.

2. Formation of a set of initial data of IS.

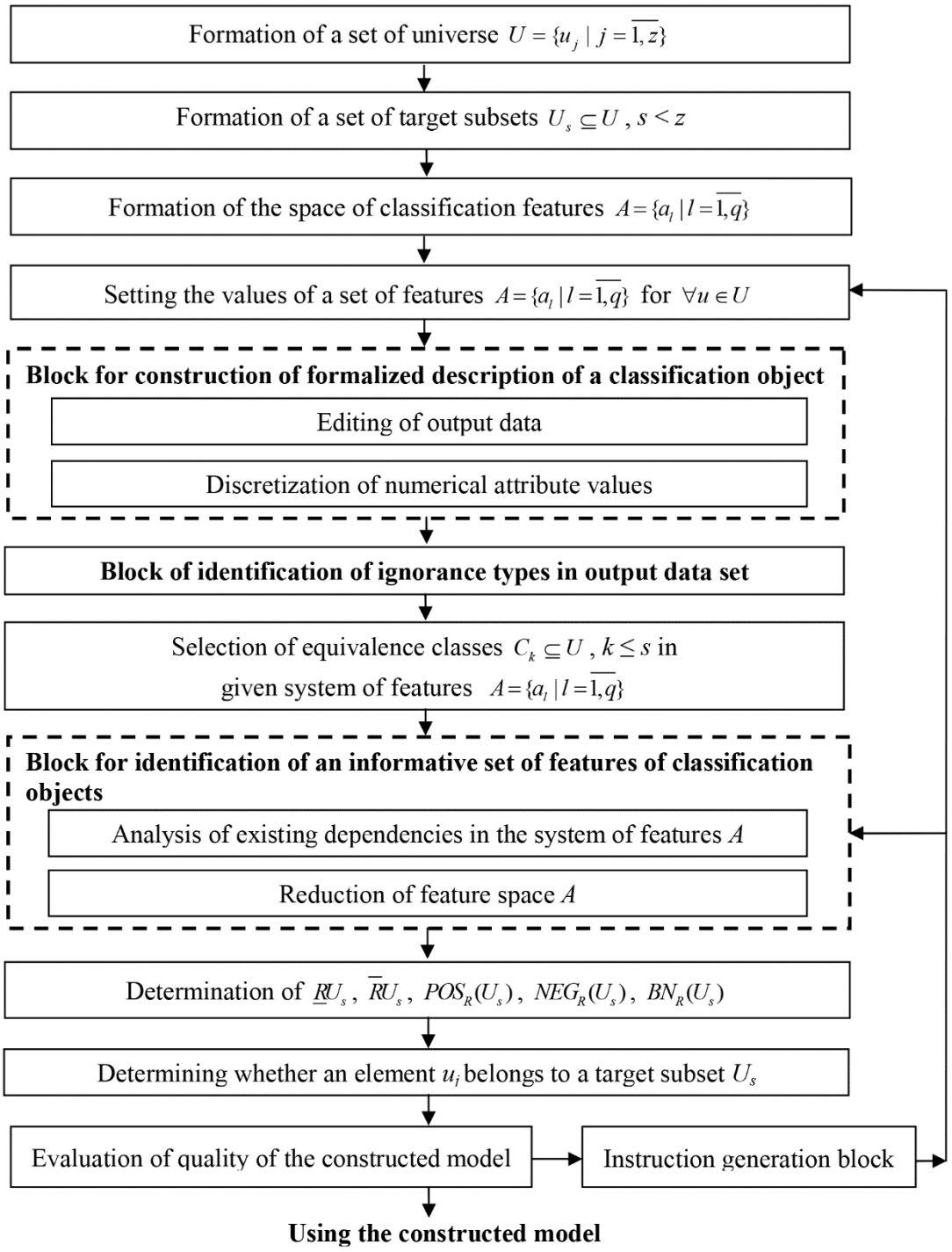

Figure 1 – Structure of IT for structuring (classification) of imprecise (rough) data and expert knowledge

At this stage, the values of a given set of classification features (attributes) are established (defined). The values of a set of attributes $A = \{a_l \mid l = \overline{1,q}\}$ can be derived

from both objective and subjective data.

The obtained knowledge system can be presented in the form of a table of $z \times q$ dimensions, the rows of which correspond to the elements $u_j \in U$, and the columns correspond to the features (attributes) $a_l \in A$ of these elements. In the cell at the intersection of the *j*-th row and the *l*-th column, the value of the *l*-th attribute for the *j*-th element is displayed.

3. Construction of formalized description of classification object.

At this stage, preliminary processing and preparation of IS data is carried out in order to prepare the initial data to a form convenient for recognition (classification).

3.1 Editing of IS data, in order to reduce the size of the training set while maintaining the efficiency of the system as a whole.

3.2 Discretization.

According to [7], discretization is the process of converting a numerical attribute into a nominal one, by representing the area of the numerical attribute with a set of its values, and processing each interval obtained as a discrete (nominal) value of the attribute.

If the value of the relevant feature is a continuous value, then for the further analysis of the IS data (definition of elementary categories) appropriate intervals of the values of the analyzed feature should be selected. For example, if it is necessary to classify a set of individuals by their age, then the intervals of age values for all elementary categories should be specified, for example, for the categories "young people", "middle-aged people", "elderly people".

Discretization is a mandatory stage in the analysis of IS, since the mathematical apparatus of RST does not provide mechanisms for processing numerical attributes. Various methods of discrediting are considered in works [2, 11, 12, 21, 26].

4. Identification of ignorance types which are presented in the set of initial data.

In [5], a number of techniques for processing of IS data under incompleteness of available knowledge have been proposed.

5. Identification of equivalence classes (main categories of knowledge) $C_k \subseteq U$,

$k \le s$ in a given system of features $A = \{a_l \mid l = \overline{1,q}\}$.

The main categories of knowledge are the categories selected for the entire set of values of the relevant feature; the elementary categories are the categories selected for each value of the relevant feature. It should be noted that within an elementary category, slight differences in the values of the relevant feature are ignored, that is, if elements with different shades of gray are included in one category, these shades are ignored and all elements belonging to this category are considered gray.

6. Determination of an informative set of features of classification objects.

6.1 Analysis of existing dependencies between attributes.

It could be possible when some primitive attributes are more important for separation (classification) than others, especially when the existing knowledge system is represented in the form of IS. In this case, the problem arises of determining the coefficients of significance or weight of individual primitive attributes in IS [23].

The importance of some primitive attribute is not an absolute concept, it is determined only by the content of a specific task. An assessment of the importance of attributes can only be performed based on an analysis of the relevant separations of the elements of universe. There are a number of techniques that allow to analyze the existing dependencies between attributes and evaluating their importance in IS [16, 20]. In [23], a technique for assessing the importance of individual attributes (or any subsets of such attributes) for separating elements of the universe by the entire set of relevant attributes has been proposed.

6.2. Reduction of attribute space.

Some IS attributes do not carry significant information, their use does not sufficiently affect the efficiency of the classification, moreover, it can lead to the deterioration of the performance of the classifier, noisy data, an unjustified increase in the amount of information used and processed, which in turn contributes to the growth of computational costs and used in the process classification of resources (time, memory). This gives rise to the task of knowledge reduction (attribute space), i.e., the process of selection of such essential part of knowledge that is sufficient to define all relevant concepts [23].

Reduction of the attribute space is a resource-intensive computational procedure that requires the search for subsets of n initial features (attributes or attribute-conditions) in the space of 2^n-1 possible subsets in accordance with a predetermined evaluation criterion.

The main components of the algorithm for selection of attributes (features) are: an evaluation function used to calculate the necessity of a subset of features; generation procedures responsible for generation of different subsets of attributes candidates.

Among the most frequently used indicators for assessing the suitability (necessity) of a subset of attributes, it is possible to single out the assessment of the quality of the approximation (classification) [3, 4], etc.; conditional independence, and approximate entropy [20].

The synthesis of IS reducts is the result of the knowledge reduction process (reduction of the attribute space). The reduct is a minimal (sufficient) set of attributes (features) that provides the same granularity of the universe as the entire original set of attributes, that is, such a set of attributes that by itself fully characterizes the existing IS knowledge.

A comparative analysis of IS attribute reduction algorithms showed that the QuickReduct algorithm is the most effective and widespread [13, 19, 24]. Among the main approaches used in the reduction of attributes, the following can be distinguished: heuristic search [8, 14, 27]; genetic algorithms, ant algorithms [3, 4, 9, 25]; hybrid approaches using elements of rough set theory and fuzzy logic [1, 6, 10]; approaches aimed at determining the dependence between attributes and the weight of the initial attributes [22].

7. Determination of $\underline{R}U_s$, $\overline{R}U_s$, approximations, $POS_R(U_s)$, $NEG_R(U_s)$, $BN_R(U_s)$ values of the target set U_s, and establishing the belonging of $u_j \in U$ to the given target set U_s.

In [16], the following scheme for establishing the fact of belonging of the object (example) under consideration to one of the specified target sets U_s is defined. For each new element and for each target set, it is checked whether this element belongs to the

positive $POS_R(U_s)$, negative $NEG_R(U_s)$ or boundary region $BN_R(U_s)$ of each target set U_s:

1. if the element $u_j \in U$ belongs to the lower approximation $\underline{RU}_s$ (positive region) of the target set U_s, then this element can definitely be identified as belonging to the target set U_s;

2. if the element $u_j \in U$ belongs to the negative region $NEG_R(U_s)$ of the target set U_s, then this element can definitely be identified as not belonging to the target set U_s;

3. if the element $u_j \in U$ belongs to the boundary region $BN_R(U_s)$ of the target set U_s, then it is impossible to say anything definite about whether or not the element u_j belongs to the target set U_s. This is the area of uncertainty, for correct operation with which RST was proposed.

8. Evaluation of the quality of constructed model.

The accuracy of the approximation can be defined as

$$\alpha_R(C) = \frac{\sum|\underline{R}X_i|}{\sum|\overline{R}X_i|}, \; i = \overline{1,n}. \tag{6}$$

The accuracy of the approximation reflects the probability of correct (unambiguous) classifications based on the existing knowledge of R.

The quality of the approximation can be defined as

$$\eta_R(C) = \frac{\sum|\underline{R}X_i|}{|U|}, \; i = \overline{1,n}. \tag{7}$$

The quality level of the model reflects the percentage of correctly classified examples (elements) based on the existing knowledge R. If the accuracy of the model is permissible (acceptable), then it is possible to use the model to classify new data.

Conclusions

The problem of structuring of imprecise (rough, unprocessed, unordered) data and expert knowledge has been considered. The tasks that arise in the process of analysis, structuring and processing of information system data are defined.

The methodology of synthesis and generalized structure of IT for structuring of imprecise (rough) data and expert knowledge in IS has been proposed. The proposed IT can be used to solve the problem of classifying rough (unprocessed, imprecise) arrays of data in the presence of such forms of ignorance as inaccuracy, inconsistency, incompleteness of the source information (data, knowledge).