ROUGH SET APPROXIMATIONS: A CONCEPT ANALYSIS POINT OF VIEW

Yiyu Yao

University of Regina, Regina, Saskatchewan, Canada

Keywords: Concept analysis, data processing and analysis, description language, form and content of data, definable concepts, lower and upper approximations, rough set approximations

Contents

Summary

Rough set theory was proposed by Pawlak for analyzing data and reasoning about data. From a concept analysis point of view, we review and reformulate main results of rough set theory in the context of data processing and analysis. This enables us to see clearly the motivations for introducing rough set theory and its basic components and appropriate applications, leading to an appreciation for the theory.

Information about this paper:
Yao, Y.Y.: Rough set approximations: A concept analysis point of view. In: Ishibuchi, H. (Ed.), Computational Intelligence - Volume I, Encyclopedia of Life Support Systems (EOLSS), pp. 282-296, 2015.

1 Two Aspects of Data

In order to see the motivations for introducing rough set theory and, hence, its uniqueness and contributions, we first give a brief discussion of two important aspects of data and then present an interpretation of rough sets as a theory concerning the meaning of data from a concept analysis point of view.

In processing and analyzing data, we consider two important aspects of data, namely, the form and content of data. Consequently, there are two fundamental classes of tasks: one is the class of form-oriented tasks and the other the class of content-oriented tasks. Form-oriented tasks focus on manipulating data as uninterpreted symbols, such as communication, storage and retrieval of data, without considering their physical meaning. Content-oriented tasks concentrate on semantics of data, such as determining the meaning of data, providing an explanation of data, building models from data etc., without worrying about how data are stored, retrieved and communicated. The division of the two (i.e., separation of form and content), on the one hand, and the union (i.e., integration of form and content), on the other hand, are crucial to data processing and analysis.

Normally, the separation of form and content leads to a simple and general theory for data processing and analysis at symbolic level. Two examples of form-oriented data processing are the information theory of communications proposed by Shannon [13] and the relational database theory proposed by Codd [1] for storing and retrieving data. Shannon's theory focuses on "reproducing at one point either exactly or approximately a message selected at another point." The meaning of the messages is considered to be irrelevant for purpose of transmitting the messages. Codd's theory is "concerned with the application of elementary relation theory to systems which provide shared access to large banks of formatted data." Data are represented conceptually by using "n-ary relations, a normal form for data base relations," for retrieval, independent of particular machine implementations and specific applications. The meaning of data in a database is not considered.

For content-oriented tasks, the semantics of data is of the main concern. We determine the meaning of data independent of the form or appearance of data as well as the methods for communicating, storing or retrieving data. Unlike the form-oriented tasks, it might be difficult to have a simple and general theory for modelling semantics of data, as semantics is usually domain and context dependent. Rough set analysis [10, 11] and formal concept analysis [2, 18] are two theories, proposed at the same time, for describing

and studying definable concepts and the structures of all definable concepts in data represented in a tabular form as in relational databases.

Concepts are the basic units of thought that underlie human intelligence and communication. A study of concepts involves multiple disciplines, including philosophy, psychology, cognitive science, mathematics, inductive data processing and analysis, inductive learning, and many others [7, 15, 16, 17]. There are many views of concepts such as the classical view, the exemplar view, the frame view, and the theory view [17]. In the classical view, concepts have well-defined boundaries and are describable by sets of singly necessary and jointly sufficient conditions [17]. Every concept consists of two parts, the intension and the extension of the concept [8, 15, 16, 17]. The intension of a concept consists of all properties or attributes that are valid for all those objects to which the concept applies. The extension of a concept is the set of objects or entities that are instances of the concept.

Due to the complexity and diversity of concepts, it is difficult to design a method that is general enough for describing intensions of all concepts. Instead, we build a specific model that enables us to define explicitly and precisely a certain class of concepts in a particular context. Formal concept analysis, proposed by Wille [18], investigates a concept that is defined by and only defined by a set of attributes in a binary data/information table called a formal context. The set of all formal concepts, i.e., all definable concepts, forms a lattice, showing the hierarchical relationships between concepts. Significant contributions of formal concept analysis are an explicit and precise description of the intension and extension of a concept, and the characterization of relationships between concepts using a lattice.

Rough set theory is another theory for concept analysis using an information table. Although earlier studies [4, 5, 6, 9] aimed at formulating a mathematical foundation of information systems characterized by information tables, the main contributions of rough set theory are the introduction of the notion of definability of concepts/sets [5, 26] and the approximations of a set by a pair of definable sets.

In this paper, we only examine the two notions of definability and approximations. For a more complete discussion on all aspects of rough set theory and its applications, a reader may read the book by Pawlak [11] and some recently edited books [12, 14]. For studies on the connections between formal concept analysis and rough set analysis, a reader may read some recent papers (for example, [3, 19, 25, 32]).

2 Definability and Approximations

Rough set analysis is based on two basic notions of the definability of concepts and the approximation of concepts. These two notions are defined with respect to an information table that describes all available information of a set of objects.

2.1 Information tables

An information table T can be defined as a tuple as follows [9, 11]:

$$T = (U, AT, \{V_a \mid a \in AT\}, \{I_a \mid a \in AT\}), \tag{1}$$

where U is a finite set of objects called the universe, AT is a finite set of attributes, V_a is the domain of attribute a, and $I_a : U \longrightarrow V_a$ is an information function. We use $I_a(x)$ to denote the value of object x on attribute a. We can conveniently represent an information table in a tabular form, in which each row represents an object, each column represents an attribute, and each cell represents the value of an object on the corresponding attribute.

Table 1 is an information table with $U = \{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}$, $AT = \{\text{Height, Hair, Eyes}\}$, $V_{\text{Height}} = \{\text{short, tall}\}$, $V_{\text{Hair}} = \{\text{blond, red, dark}\}$ and $V_{\text{Eyes}} = \{\text{blue, brown}\}$. For object o_1, we have:

$$\begin{aligned} I_{\text{Height}}(o_1) &= \text{short}, \\ I_{\text{Hair}}(o_1) &= \text{blond}, \\ I_{\text{Eyes}}(o_1) &= \text{blue}. \end{aligned}$$

In a table representation, objects are given in a sequence of rows and attributes are in a sequence of columns. Although in the literature of rough sets one typically refers to an object by its row number or an attribute by its column number, it is important to note that semantically there is no ordering on the set of objects nor on the set of attributes. From the table, it can be seen that some objects have the same description. For example, objects o_2 and o_3 have the same description. Consequently, based only on their description, one can not distinguish objects o_2 and o_3. This observation is in fact the basis of rough set analysis.

Object	Height	Hair	Eyes
o_1	short	blond	blue
o_2	short	blond	brown
o_3	short	blond	brown
o_4	tall	dark	blue
o_5	tall	dark	blue
o_6	tall	dark	blue
o_7	tall	red	blue

Table 1: An information table

2.2 Concepts and definable concepts

In an information table, a subset of objects $X \subseteq U$ may be viewed as the extension of a concept. In order to describe formally the intension of a concept, we introduce a description language, as suggested by Marek and Pawlak [5]. A description language DL can be recursively defined as follows:

$$\begin{aligned} &(1) \qquad (a = v) \in DL, \text{ where } a \in AT, v \in V_a, \\ &(2) \qquad \text{if } p, q \in DL, \text{ then } (p \wedge q), (p \vee q) \in DL. \end{aligned}$$

Formulas defined by (1) are called atomic formulas. For simplicity, we consider a language defined by two logic connectives $\wedge$ and $\vee$, which is a sublanguage used by Marek and Pawlak [5] and by Pawlak [11]. By assuming that $\wedge$ has a higher precedence in computation, one may remove unnecessary parentheses in a formula. This language is powerful enough for rough set analysis.

The satisfiability of a formula p by an object x, written $x \models p$, is defined as follows:

$$\begin{aligned} &\text{(i)} \qquad x \models a = v, \text{iff } I_a(x) = v, \\ &\text{(ii)} \qquad x \models p \wedge q, \text{iff } x \models p \text{ and } x \models q, \\ &\text{(iii)} \qquad x \models p \vee q, \text{iff } x \models p \text{ or } x \models q. \end{aligned}$$

If p is a formula, the set $m(p) \subseteq U$ defined by:

$$m(p) = \{x \in U \mid x \models p\} \tag{2}$$

is called the meaning set of formula p. The meaning set $m(p)$ consists of all objects that satisfy p. With the introduction of meaning sets, we can establish the following linkage between logic and set operations:

$$\begin{aligned} &\text{(m1)} && m(a = v) = \{x \in U \mid I_a(x) = v\}, \\ &\text{(m2)} && m(p \wedge q) = m(p) \cap m(q), \\ &\text{(m3)} && m(p \vee q) = m(p) \cup m(q). \end{aligned} \tag{3}$$

That is, logic conjunction and disjunction are interpreted in terms of set intersection and union, respectively.

In Table 1, an example of atomic formulas is Height = tall, indicating that the Height of an object is tall. By definition, Hair = blond$\wedge$Eyes = brown is a formula, indicating that the Hair of an object is blond and, at the same time, the Eyes of the object is brown. As examples to demonstrate satisfiability, we have $o_1 \models$ Height = short and $o_3 \models$ Hair = blond $\wedge$ Eyes = brown. A few examples for computing the meaning sets are given by:

$$\begin{aligned} &m(\text{Height} = \text{short}) = \{o_1, o_2, o_3\}, \\ &m(\text{Hair} = \text{dark}) = \{o_4, o_5, o_6\}, \\ &m(\text{Height} = \text{short} \wedge \text{Hair} = \text{dark}) = \{o_1, o_2, o_3\} \cap \{o_4, o_5, o_6\} = \emptyset, \\ &m(\text{Height} = \text{short} \vee \text{Hair} = \text{dark}) = \{o_1, o_2, o_3\} \cup \{o_4, o_5, o_6\} = \\ &\quad \{o_1, o_2, o_3, o_4, o_5, o_6\}. \end{aligned}$$

According to the first equation, Height = short may be viewed as an intension of a concept whose extension is $\{o_1, o_2, o_3\}$.

The meaning-set function m maps a formula p to a unique subset $m(p)$ of U. On the other hand, the reverse process is not so simple. For a subset $X \subseteq U$, we may or may not find a formula p such that $X = m(p)$. For the former case, there may also exist more than one formula. For example, we can not found a formula to describe the set of objects $\{o_1, o_2\}$, because any formula that describes o_2 also describes o_3 and it is impossible to separate o_2 and o_3. For the subset of objects $\{o_1, o_2, o_3\}$, we can use Height = short or Hair = blond to describe it. In fact, this is main difference between rough set analysis and formal concept analysis. For the latter, it is required that one must find a unique intension of a concept.

Suppose a set of objects $X \subseteq U$ is the extension of a concept. We call X a definable concept or a definable set if there exists a formula p such that

$$X = m(p), \tag{4}$$

otherwise, X is an undefinable set. From a formula, we form a definable concept $(p, m(p))$. In rough set analysis, the notion of definability is based on a one-way definability. That is, p defines $m(p)$. It may also happen that there exist another formula q that is different from p and $m(p) = m(q)$. That is, the extension of a definable concept does not uniquely determine an intension of a concept.

Let $\mathrm{DEF}(U)$ denote the set of all definable sets. It can be verified that, given a finite set of objects and a finite set of attributes, $\mathrm{DEF}(U)$ contains the empty set $\emptyset$, the universe U, and is closed under set complement, intersection and union. In other words, $\mathrm{DEF}(U)$ is a sub-Boolean algebra of 2^U defined by the power set of U. Any set in $2^U - \mathrm{DEF}(U)$ is an undefinable set.

Consider Table 1, the family of all definable sets are given by

$$\begin{aligned}\mathrm{DEF}(U) \;=\; \{&\emptyset,\\ &\{o_1\}, \{o_2, o_3\}, \{o_4, o_5, o_6\}, \{o_7\},\\ &\{o_1, o_2, o_3\}, \{o_1, o_4, o_5, o_6\}, \{o_1, o_7\},\\ &\{o_2, o_3, o_4, o_5, o_6\}, \{o_2, o_3, o_7\}, \{o_4, o_5, o_6, o_7\}\\ &\{o_1, o_2, o_3, o_4, o_5, o_6\}, \{o_1, o_2, o_3, o_7\}\\ &\{o_1, o_4, o_5, o_6, o_7\}, \{o_2, o_3, o_4, o_5, o_6, o_7\},\\ &U\}.\end{aligned}$$

Figure 1 is the Boolean algebra of all definable sets, where a set $\{o_2, o_3\}$ is simply represented by $\{2, 3\}$. The power 2^U contains $2^7 = 128$ subsets, of which only $2^4 = 16$ are definable.

2.3 Approximations of concepts

In an information table, some concepts are definable while others are undefinable. To make inference about an undefinable concept, we must approximate it by using definable concepts.

For a subset $X \subseteq U$, we can approximate it by a pair of lower and upper approximations:

$$\begin{aligned}\underline{apr}(X) &= \text{the greatest definable set contained by } X,\\ \overline{apr}(X) &= \text{the least definable set containing } X.\end{aligned} \tag{5}$$

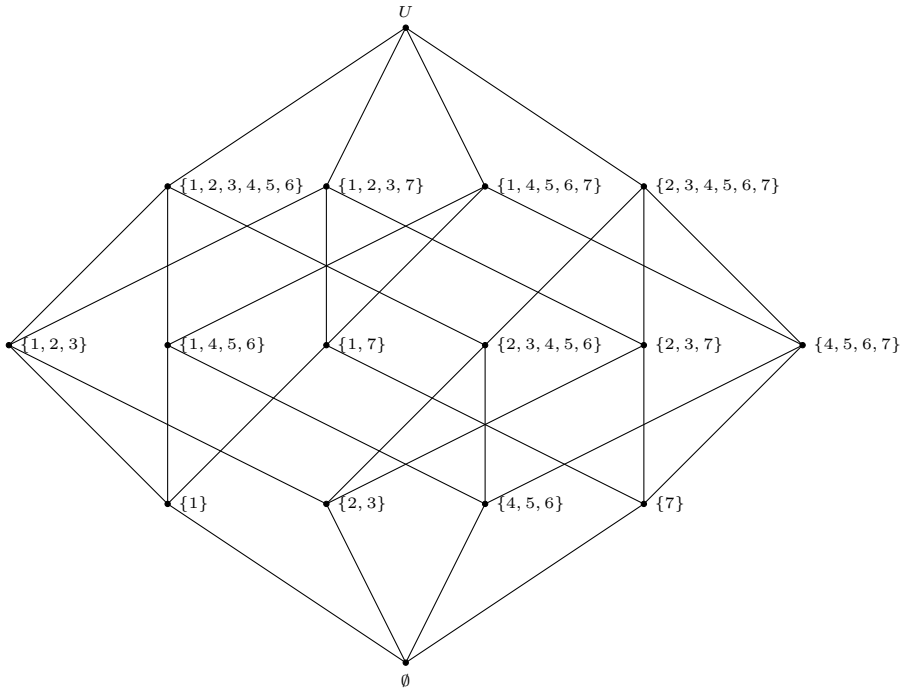

Figure 1: The family of definable sets

By definition, the following properties hold:

$$
\begin{aligned}
&\text{(I)} && \underline{apr}(X) \subseteq X \subseteq \overline{apr}(X),\\
&\text{(II)} && X \in \mathrm{DEF}(U) \Longleftrightarrow \underline{apr}(X) = X = \overline{apr}(X),\\
&\text{(III)} && X \subseteq Y \Longrightarrow \underline{apr}(X) \subseteq \underline{apr}(Y),\\
& && X \subseteq Y \Longrightarrow \overline{apr}(X) \subseteq \overline{apr}(Y),\\
&\text{(IV)} && \underline{apr}(X) = (\overline{apr}(X^c))^c,\\
& && \overline{apr}(X) = (\underline{apr}(X^c))^c,\\
&\text{(V)} && \underline{apr}(X \cap Y) = \underline{apr}(X) \cap \underline{apr}(Y)\\
& && \overline{apr}(X \cup Y) = \overline{apr}(X) \cup \overline{apr}(Y),
\end{aligned}
$$

where $(\cdot)^c$ denotes the complement of a set. Property (I) indicates that a set is approximated from below and above by two definable sets. Property (II) is particularly interesting, as it shows that definable sets are in fact approximated by themselves. Property (III) states that both approximations are

monotonic with respect to the set-inclusion relation. According to property (IV), the two approximations may be viewed as dual approximations and we only need to define one of them. Property (V) shows the possible construction rules for building the approximations of the intersection or the union of two sets from their approximations.

An alternative approach to define and interpret rough set approximations is based on three regions. For a subset $X \subseteq U$, we can divide the universe U into three pair-wise disjoint regions, namely, the positive, negative and boundary regions, as follows:

$$\begin{aligned} \mathrm{POS}(X) &= \text{the greatest definable set contained by X,} \\ \mathrm{NEG}(X) &= \text{the greatest definable set contained by X}^{\mathrm{c}}, \\ \mathrm{BND}(X) &= (\mathrm{POS}(X) \cup \mathrm{NEG}(X))^c. \end{aligned} \tag{6}$$

By definition, all three regions are definable sets. In this way, a set is approximated by three pair-wise disjoint definable sets, and some of them can be the empty set $\emptyset$.

Figure 2 illustrates the rough set approximation of a set. The figure has a geometrically and intuitively appealing interpretation of rough set approximations. An equivalence class is represented by a small square which is a smallest nonempty definable set. A definable set may be viewed as a "regular" shape defined by a family of equivalence classes and may be arranged into a group of rectangles in the figure; an undefinable set may be viewed as an "irregular" shape. Rough sets are viewed as approximations of irregular shapes by regular shapes, a common practice in geometry.

The pair of lower and upper approximations and the three regions are two different forms, but mathematically equivalent, of rough set approximations. They determine each other as follows:

$$\begin{aligned} \mathrm{POS}(X) &= \underline{apr}(X), \\ \mathrm{NEG}(X) &= \underline{apr}(X^c) = (\overline{apr}(X))^c, \\ \mathrm{BND}(X) &= \overline{apr}(X) - \underline{apr}(X), \end{aligned} \tag{7}$$

and

$$\begin{aligned} \underline{apr}(X) &= \mathrm{POS}(X), \\ \overline{apr}(X) &= \mathrm{POS}(X) \cup \mathrm{BND}(X). \end{aligned} \tag{8}$$

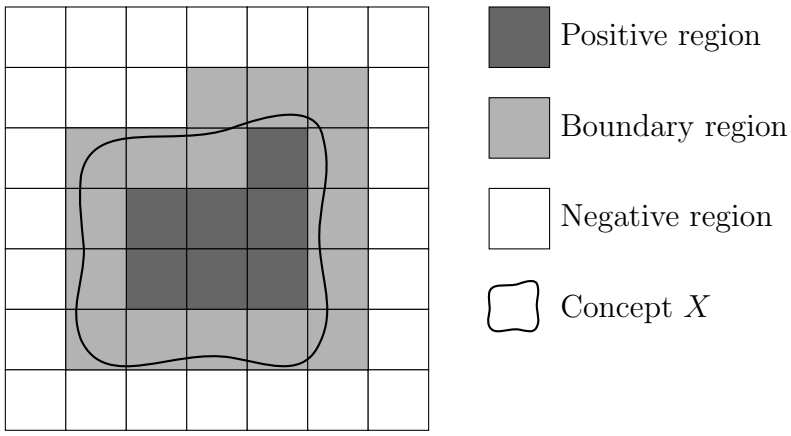

Figure 2: Approximation of a set by three regions

One may build a theory of rough sets either based on the pair of approximations or three pair-wise disjoint regions. Each formulation has its advantages. The pair of approximations explicitly gives the range within which lies the set. The three regions are closely related to a theory of three-way decisions [27, 28, 29, 30].

3 Construction of Approximations

In the definitions of rough set approximations presented in the last section, definability of sets is a basic notion. Rough set approximations, either as a pair of definable sets or as three regions, are considered as a derived notion from definability. The formulation shows the motivations for introducing approximations and is semantically meaningful. On the other hand, it is difficult to construct approximations directly from the definitions. In developing rough set theory, one typically defines approximations based on an equivalence relation by assuming that a union of a family of equivalence classes is a definable set. Although both formulations are mathematically equivalent, the equivalence relation based approach suffers from a semantic difficulty.

In the paper, we present a formulation of rough set theory consisting of two models. A semantically sound model for defining rough set approximations, which is given in the last section, and a computationally efficient model for constructing approximations, which is given in this section.

3.1 Definable sets and the Boolean algebra induced by an equivalence relation

A binary relation on U is called an equivalence relation if it is reflexive, symmetric and transitive. In an information table, one can construct equivalence relations by using a single attribute or a subset of attributes. For an attribute $a \in AT$, we can define an equivalence relation E_a as follows: for $x, y \in U$,

$$xE_a y \Longleftrightarrow I_a(x) = I_a(y). \tag{9}$$

The equivalence class containing x is denoted by $[x]_a$. Similarly, for a subset of attributes $A \subseteq AT$, we define an equivalence relation E_A:

$$xE_A y \Longleftrightarrow \forall a \in A(I_a(x) = I_a(y)). \tag{10}$$

The equivalence class containing x is denoted by $[x]_A$. By definition, it follows that, for $a \in AT$ and $A \subseteq AT$,

$$\begin{aligned} &E_{\{a\}} = E_a, \quad [x]_{\{a\}} = [x]_a, \\ &E_A = \bigcap_{a \in A} E_a, \quad [x]_A = \bigcap [x]_a. \end{aligned} \tag{11}$$

That is, we can construct the equivalence relation induced by a subset of attributes A by using equivalence relations induced by individual attributes in A.

Consider the equivalence relation E_A induced by a subset of attributes $A \subseteq AT$. The equivalence relation E_A induces a partition U/E_A of U, i.e., a family of nonempty and pair-wise disjoint subsets whose union is the universe. For an object $x \in U$, its equivalence class is given by:

$$[x]_A = \{y \in U \mid xE_A y\}. \tag{12}$$

One can construct an atomic sub-Boolean algebra $B(U/E_A)$ of 2^U with U/E_A as the set of atoms:

$$B(U/E_A) = \{\bigcup F \mid F \subseteq U/E_A\}. \tag{13}$$

Each element in $B(U/E_A)$ is the union of a family of equivalence classes. The Boolean algebra $B(U/E_A)$ contains the empty set $\emptyset$, the whole set U, and is closed under set complement, intersection, and union. The three notions

of an equivalence relation E_A, the partition U/E_A and the atomic Boolean algebra $B(U/E_A)$ uniquely determine each other. We can therefore use E_A, U/E_A and $B(U/E_A)$ interchangeably.

For Table 1, the partition of the equivalence relation E_{AT} is given by:

$$U/E_{AT} = \{\{o_1\}, \{o_2, o_3\}, \{o_4, o_5, o_6\}, \{o_7\}\}.$$

The atomic Boolean algebra is also given by Figure 1. Equivalence classes (i.e., atoms of the Boolean algebra $B(U/E_{AT})$) are the smallest nonempty definable sets. The atomic Boolean algebra $B(U/E_{AT})$ is the same as the family of all definable sets $\mathrm{DEF}(U)$. As demonstrated next, these observations are not coincident; they are true in general.

For a subset of attributes $A \subseteq AT$, if we restrict the formulas of DL by using only attributes in A, we obtain a sub-language $DL(A) \subseteq DL$. Let $\mathrm{DEF}_A(U)$ denote the family of all definable sets defined by the language $DL(A)$. It can be proved that the family of all definable sets $\mathrm{DEF}_A(U)$ is exactly the Boolean algebra $B(U/E_A)$.

With respect to a subset of attributes $A \subseteq AT$, each object x is described by a logic formula,

$$\bigwedge_{a \in A} a = I_a(x), \tag{14}$$

where $I_a(x) \in V_a$ and the atomic formula $a = I_a(x)$ indicates that the value of an object on attribute a is $I_a(x)$. The equivalence class containing x, namely, $[x]_{E_A}$, is the set of those objects that satisfy the formula $\bigwedge_{a \in A} a = I_a(x)$. In this case, we have:

$$m(\bigwedge_{a \in A} a = I_a(x)) = [x]_{E_A}. \tag{15}$$

That is, $[x]_{E_A}$ is a definable set. The formula $\bigwedge_{a \in A} a = I_a(x)$ is a description of objects that are equivalent to x with respect to A, including x itself. From the definability of equivalence classes and the relationship between logic disjunction $\vee$ and set union $\cup$, i.e., (m3) in Equation (3), it follows that the union of a family of equivalence classes is a definable set. Thus, any set in the atomic Boolean algebra $B(U/E_A)$ is a definable set. Conversely, we can show that any definable set in $\mathrm{DEF}_A(U)$ is a member of the Boolean algebra of $B(U/E_A)$ by transforming a formula into its disjunction-of-conjunction normal form.

3.2 New constructive definitions of approximations

Based on the equivalence of $\mathrm{DEF}_A(U)$ and $B(U/E_A)$, we can equivalently define rough set approximations by using the equivalence classes $[x]_A$ or the atomic Boolean algebra $B(U/E_A)$. For simplicity, we also simply write $[x]_A$ as $[x]$ and E_A as E when no confusion arises.

Yao [24] classifies commonly used definitions of rough set approximations into three types: the element-based definition, the granule-based definition and the subsystem-based definition. They are given by: for $X \subseteq U$,

- Element-based definition:

$$\begin{aligned}
&\underline{apr}(X) = \{x \in U \mid [x] \subseteq X\},\\
&\overline{apr}(X) = \{x \in U \mid [x] \cap X \neq \emptyset\};\\
\\
&\mathrm{POS}(X) = \{x \in U \mid [x] \subseteq X\},\\
&\mathrm{NEG}(X) = \{x \in U \mid [x] \cap X = \emptyset\},\\
&\mathrm{BND}(X) = \{x \in U \mid ([x] \not\subseteq X \wedge ([x] \cap X \neq \emptyset)\}.
\end{aligned} \tag{16}$$

- Granule-based definition:

$$\begin{aligned}
&\underline{apr}(X) = \bigcup\{[x] \in U/E \mid [x] \subseteq X\},\\
&\overline{apr}(X) = \bigcup\{[x] \in U/E \mid [x] \cap X \neq \emptyset\};\\
\\
&\mathrm{POS}(X) = \bigcup\{[x] \in U/E \mid [x] \subseteq X\},\\
&\mathrm{NEG}(X) = \bigcup\{[x] \in U/E \mid [x] \cap X = \emptyset\},\\
&\mathrm{BND}(X) = \bigcup\{[x] \in U/E \mid ([x] \not\subseteq X \wedge ([x] \cap X \neq \emptyset)\}.
\end{aligned} \tag{17}$$

- Subsystem-based definition:

$$\begin{aligned}
&\underline{apr}(X) = \bigcup\{K \in B(U/E) \mid K \subseteq X\},\\
&\overline{apr}(X) = \bigcap\{K \in B(U/E) \mid X \subseteq K\};\\
\\
&\mathrm{POS}(X) = \bigcup\{K \in B(U/E) \mid K \subseteq X\},\\
&\mathrm{NEG}(X) = \bigcup\{K \in B(U/E) \mid K \subseteq X^c\},\\
&\mathrm{BND}(X) = (\mathrm{POS}(X) \cup \mathrm{NEG}(X))^c.
\end{aligned} \tag{18}$$

The subsystem-based definition is in fact the same as the definition given in the last section.

Although the three definitions are mathematically equivalent, they offer different hints when we generalize Pawlak rough sets. The element-based definition enables us to establish a connection between rough sets and modal logics, offering a direction for generalizing rough sets by using non-equivalence relations [20, 21]. The granule-based definition connects rough sets and granular computing [23], offering another direction for generalizing rough sets by using coverings [21]. The subsystem-based definition can be used to generalize rough sets by using other mathematical structures such as Boolean algebra, lattices, topological spaces, closure systems, and posets [22, 31].

4 Conclusion

Form and content are two important aspects of data. Rough set analysis focuses on the content, i.e., meaning, aspect of data. From a concept analysis point of view, we present two formulations/models of rough set approximations. One formulation is based on the notion of the definability of concepts, which enables us to see clearly the motivations for introducing rough set theory and to provide a semantically sound interpretation of rough set approximations. The other formulation is based on an equivalence relation, which leads to a computationally efficient method for constructing approximations. The integration of the two models results in a full understanding of rough set approximations.

Glossary

classical view of concepts: Concepts have well-defined boundaries. Every concept is described by its intension (i.e., all properties or attributes that are valid for all those objects to which the concept applies) and extension (i.e., the set of objects that are instances of the concept).

definable sets: A set is definable if there exists a formula such that the set is exactly the set of objects satisfying the formula.

description language: A language used to describe a set of objects by using a set of attributes.

form and content of data: form refers to the symbolic representation of data and content refers to the meaning of data.

lower and upper approximations: The lower approximation of a set is the greatest definable set contained in the set. The upper approximation of a set is the least definable set containing the set.

rough set three regions: The positive region of a set is the greatest definable set contained in the set. The negative region of a set the greatest definable set contained in the complement of the set. The boundary region is the complement of the union of the positive and negative regions.

Bibliography

[1] Codd, E.F., A relational model of data for large shared data banks, Communications of the ACM, 13, 377-387, 1970. [This paper introduces the theory of relational databases.]

[2] Ganter, B., Wille, R., Formal Concept Analysis: Mathematical Foundations, Springer, New York, 1999. [This is a seminal book on the theory and applications of formal concept analysis by its inventor (second author).]

[3] Lai, H.L., Zhang, D.X., Concept lattices of fuzzy contexts: Formal concept analysis vs. rough set theory, International Journal of Approximate Reasoning, 50, 695-707, 2012. [Discusses connections of formal concept analysis and rough set theory with reference to a fuzzy formal context.]

[4] Marek, V.W., Zdzisław Pawlak, databases and rough sets, in: Skowron, A., Suraj, Z. (eds.), Rough Sets and Intelligent Systems, Springer, Berlin, 175-184, 2013. [This paper recalls several events that had led to the introduction of rough set theory. It discusses motivations for rough set theory.]

[5] Marek, V.W., Pawlak, Z., Information storage and retrieval systems: Mathematical foundations, Theoretical Computer Science, 1, 331-354, 1976. [This paper discusses the use of a description language. The language used in present paper is only a sublanguage.]

[6] Marek, V.W., Truszczyński, M., Contributions to the theory of rough sets, Fundamenta Informaticae, 39, 389-409, 1999. [This paper examines the contributions of rough set theory.]

[7] Michalski, R.S., Carbonell, J.G., Mitchell, T.M. (eds.), Machine Learning, an Artificial Intelligence Approach, Morgan Kaufmann Publishers, Inc., Los Altos, California, 1983. [An edited book on many topics in machine learning.]

[8] Ogden, C.K., Richards I.A. The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism, 8th edition, Harcourt Brace, New York, 1946. [The notion of meaning triangle introduced in this book serves as a basis for understanding the classical view of concepts.]

[9] Pawlak, Z., Information systems - theoretical foundations, Information Systems, 6, 205-218, 1981. [This paper discusses the notion of information systems, which is called information tables in the present paper.]

[10] Pawlak, Z., Rough sets, International Journal of Computer and Information Sciences, 11, 341-356, 1982. [This paper introduces rough set theory.]

[11] Pawlak, Z., Rough Sets, Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht, 1991. [A seminal book on the theory of rough sets by its inventor.]

[12] Peters, G., Lingras, P., Ślęzak, D., Yao, Y.Y. (eds.), Rough Sets: Selected Methods and Applications in Management and Engineering, Springer, London, 2012. [A collection of papers on business and engineering applications of rough sets.]

[13] Shannon, C.E., A mathematical theory of communication, The Bell System Technical Journal, 27, 379-423, 623-656, 1948. [The two papers introduce information theory.]

[14] Skowron, A., Suraj, Z. (eds.), Rough Sets and Intelligent Systems - Professor Zdzisław Pawlak in Memoriam, Volumes 1 and 2, Springer, Berlin, 2013. [The two volumes, in memoriam of Professor Pawlak, are a collection of articles by experts of rough sets.]

[15] Smith, E.E., Concepts and induction, in: M.I. Posner (ed.), Foundations of Cognitive Science, The MIT Press, Cambridge, Massachusetts, 501-526, 1989. [Discusses many fundamental issues of concepts and induction.]

[16] Sowa, J.F., Conceptual Structures, Information Processing in Mind and Machine, Addison-Wesley, Reading, Massachusetts, 1984. [The book introduces conceptual structures for modelling information processing.]

[17] van Mechelen, I., Hampton, J., Michalski, R.S., Theuns, P. (eds.), Categories and Concepts: Theoretical Views and Inductive Data Analysis, Academic Press, New York, 1993. [A collection of articles on categories and concepts in the context of data analysis.]

[18] Wille, R., Restructuring lattice theory: An approach based on hierarchies of concepts, in: Rival, I. (Ed.), Ordered Sets, Reidel, Dordrecht, 445-470, 1982. [The article introduces formal concept analysis.]

[19] Wolff, K.E., A conceptual view of knowledge bases in rough set theory, Ziarko, W., Yao, Y.Y. (eds.), RSCTC 2000, LNCS (LNAI) 2005, Springer, Heidelberg, 220-228, 2001. [This paper discusses a conceptual view for comparing and unifying formal concept analysis and rough set analysis.]

[20] Yao, Y.Y., Two views of the theory of rough sets in finite universes, International Journal of Approximate Reasoning, 15, 291-317, 1996. [Introduces two views for interpreting rough sets, namely, a set-oriented view and an operator-oriented view.]

[21] Yao, Y.Y., Relational interpretations of neighborhood operators and rough set approximation operators, Information Sciences, 101, 239-259, 1998. [Presents a systematic study on generalizations of rough sets by using arbitrary binary relations and coverings induced by a binary relation.]

[22] Yao, Y.Y., On generalizing Pawlak approximation operators, Polkowski, L., Skowron, A. (eds.), RSCTC 1998, LNCS (LNAI) 1424, Springer, Heidelberg, 298-307, 1998. [Investigates subsystem based generalizations of rough sets.]

[23] Yao, Y.Y., Information granulation and rough set approximation, International Journal of Intelligent Systems, 16, 87-104, 2001. [Examines rough set theory in the light of granular computing.]

[24] Yao, Y.Y., On generalizing rough set theory, Wang, G.Y., Liu, Q., Yao, Y.Y., Skowron, A. (eds.), RSFDGrC 2003, LNCS (LNAI) 2639, Springer, Heidelberg, 44-51, 2003. [Introduces three directions in generalizing rough sets, namely, generalizations by using a) an arbitrary binary relation, b) a covering of the universe, and c) a subsystem of the power set of the universe.]

[25] Yao, Y.Y., A comparative study of formal concept analysis and rough set theory in data analysis, Tsumoto, S., Słowiński, R., Komorowski, J., Grzymala-Busse, J.W. (eds.), LNCS (LNAI) 3066, Springer, Heidelberg, 59-68, 2004. [Compares rough set analysis and formal concept analysis based on the notions one-way and two-way classification rules.]

[26] Yao, Y.Y., A note on definability and approximations, LNCS Transactions on Rough Sets, VII, LNCS 4400, 274-282, 2007. [Reformulates rough sets based on the notion of definability.]

[27] Yao, Y.Y., Three-way decision: An interpretation of rules in rough set theory, Wen, P., Li, Y.F., Polkowski, L., Yao, Y.Y., Tsumoto, S., Wang, G.Y. (eds.), RSKT 2009, LNCS (LNAI) 5589, Springer, Heidelberg, 642-649, 2009. [This is the first paper on interpreting rough set three regions in terms of three-way decisions.]

[28] Yao, Y.Y., Three-way decisions with probabilistic rough sets, Information Sciences, 180, 341-353, 2010. [A detailed analysis of three-way decisions using probabilistic rough sets.]

[29] Yao, Y.Y., The superiority of three-way decisions in probabilistic rough set models, Information Sciences, 181, 1080-1096, 2011. [Proves that, under certain conditions, three-way decisions are superior to binary decisions and Pawlak three-way decisions.]

[30] Yao, Y.Y., An outline of a theory of three-way decisions, Yao, J.T., Yang, Y., Słowiński, R., Greco, S., Li, H.X., Mitra, S., Polkowski, L. (eds.), LNCS (LNAI) 7413, Springer, Heidelberg, 1-17, 2012. [Introduces and gives an outline of a theory of three-way decisions.]

[31] Yao, Y.Y., Chen, Y.H., Subsystem based generalizations of rough set approximations, Hacid, M.S., Murray, N.V., Raś, Z.W., Tsumoto, S. (eds.), ISMIS 2005, LNCS 3488, Springer, Heidelberg, 210-218, 2005. [Discusses subsystem based generalizations of rough sets.]

[32] Ytow, N., Morse, D.R., Roberts, D.M., Rough set approximation as formal concept, Journal of Advanced Computational Intelligence and Intelligent Informatics, 10, 606-611, 2006. [Discusses connections of rough sets and formal concept analysis.]

Biographical Sketches

Yiyu Yao is a professor of computer science with the Department of Computer Science, University of Regina, Regina, Saskatchewan, Canada. His research interests include information retrieval, three-way decisions, rough sets, fuzzy sets, interval sets, granular computing, Web intelligence, and data mining. His publications cover various topics on a triarchic theory of granular computing, a theory of three-way decisions, the foundations of data mining, modelling information retrieval systems, information retrieval support systems, generalized rough sets and many more.