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.

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 (1948) and the relational database theory proposed by Codd (1970) 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 modeling semantics of data, as semantics is usually domain and context dependent. Rough set analysis (Pawlak, 1982. 1991) and formal concept analysis (Ganter and Wille, 1999; Wille, 1982) 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 (Michalski et al, 1983; Smith, 1989; Sowa, 1984; van Mechelen et al, 1993). There are many views of concepts such as the classical view, the exemplar view, the frame view, and the theory view (van Mechelen et al, 1993). In the classical view, concepts have well-defined boundaries and are describable by sets of singly necessary and jointly sufficient conditions (van Mechelen et al, 1993). Every concept consists of two parts, the intension and the extension of the concept (Ogden and Richards, 1946; Smith, 1989; Sowa, 1984; van Mechelen et al, 1993). 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 (1982), 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 (Marek 2013; Marek and Pawlak, 1976; Marek and Truszczyński, 1999; Pawlak, 1981) 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 (Marek and Pawlak, 1976; Yao, 2007) and the approximations of a set by a pair of definable sets.

In this chapter, 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 (1991) and some recently edited books (Peters et al, 2012; Skowron and Suraj, 2013) . For studies on the connections between formal concept analysis and rough set analysis, a reader may read some recent papers (for example, Lai and Zhang, 2012; Wolff, 2001; Yao, 2004; Ytow et al, 2006).

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 (Pawlak, 1981, 1991):

$$T = \left(U, AT, \left\{V_a \middle| a \in AT\right\}, \left\{I_a \middle| a \in AT\right\}\right) \qquad (1)$$

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 $I_a : U \to V_a$ 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 $U = \{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}$, $AT = \{Height, Hair, Eyes\}$, $V_{\text{Height}} = \{short, tall\}$, $V_{\text{Hair}} = \{blond, red, dark\}$ and $V_{\text{Eyes}} = \{blue, brown\}$. For object o_1, we have:

$I_{\text{Height}}(o_1) = short$,

$I_{\text{Hair}}(o_1) = blond$,

$I_{\text{Eyes}}(o_1) = blue$.

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 (1976). A description language DL can be recursively defined as follows:

(1) $(a = v) \in DL$, where $a \in AT$, $v \in V_a$,

(2) if $p, q \in DL$, then $(p \wedge q), (p \vee q) \in DL$.

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 sub-language used by Marek and Pawlak (1976) and by Pawlak (1991). 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:

(i) $x \models a = v$, iff $I_a(x) = v$,

(ii) $x \models p \wedge q$, iff $x \models p$ and $x \models q$

(iii) $x \models p \vee q$, iff $x \models p$ or $x \models q$

-

-

-

Bibliography

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

[2] Ganter, B., Wille, R., (1999) *Formal Concept Analysis: Mathematical Foundations*, Springer, New York. [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., (2012) Concept lattices of fuzzy contexts: Formal concept analysis vs. rough set theory, *International Journal of Approximate Reasoning*, 50, 695-707. [Discusses connections of formal concept analysis and rough set theory with reference to a fuzzy formal context.]

[4] Marek, V.W., (2013) Zdzisław Pawlak, databases and rough sets, in: Skowron, A., Suraj, Z. (eds.), *Rough Sets and Intelligent Systems*, Springer, Berlin, 175-184. [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., (1976) Information storage and retrieval systems: Mathematical foundations, *Theoretical Computer Science*, 1, 331-354. [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., (1999) Contributions to the theory of rough sets, *Fundamenta Informaticae*, 39, 389-409. [This paper examines the contributions of rough set theory.]

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

[8] Ogden, C.K., Richards I.A. (1946) *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. [The notion of meaning triangle introduced in this book serves as a basis for understanding the classical view of concepts.]

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

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

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

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

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

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

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

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

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

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

[19] Wolff, K.E., (2001) *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. [This paper discusses a conceptual view for comparing and unifying formal concept analysis and rough set analysis.]

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

[21] Yao, Y.Y., (1998a) Relational interpretations of neighborhood operators and rough set approximation operators, *Information Sciences*, 101, 239-259. [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., (1998b) *On generalizing Pawlak approximation operators*, Polkowski, L., Skowron, A. (eds.), RSCTC 1998, LNCS (LNAI) 1424, Springer, Heidelberg, 298-307. [Investigates subsystem based generalizations of rough sets.]

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

[24] Yao, Y.Y., (2003) *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. [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., (2004) *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.), RSCTC 2004, LNCS (LNAI) 3066, Springer, Heidelberg, 59-68. [Compares rough set analysis and formal concept analysis based on the notions one-way and two-way classification rules.]

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

[27] Yao, Y.Y., (2009) *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. [This is the first paper on interpreting rough set three regions in terms of three-way decisions.]

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

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

[30] Yao, Y.Y., (2012) *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.), RSCTC 2012,LNCS (LNAI) 7413, Springer, Heidelberg, 1-17. [Introduces and gives an outline of a theory of three-way decisions.]

[31] Yao, Y.Y., Chen, Y.H., (2005) *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. [Discusses subsystem based generalizations of rough sets.]

[32] Ytow, N., Morse, D.R., Roberts, D.M., (2006) Rough set approximation as formal concept, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 10, 606-611. [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, modeling information retrieval systems, information retrieval support systems, generalized rough sets and many more.