A ROUGH SET BASED ALGEBRAIC APPROACH TO MODELLING COMPLEX SYSTEMS

D. SITNIKOV[1], O. RYABOV[2], I. MISHCHERIAKOV[1] & A. KOVALENKO[1]
[1] Kharkiv National University of Radio Electronics, Ukraine.
[2] National Institute of Advanced Industrial Science and Technology, Japan.

ABSTRACT

“A complex system is considered as an algebraic structure having specific properties that do not allow expressing precisely the meaning of information objects included in the system. An algebraic definition of complexity has been given. The complexity has been considered from two viewpoints. A system can be considered to be a complex one if (a) boundary regions for system’s objects and processes are broad; (b) possibilities for system’s decompositions are limited. Difficulties associated with complex system description and decomposition have been discussed in the framework of rough set methodology. A method for extracting salient features of information objects available in the system has been outlined. Some theoretical aspects of rough set based analysis of complex systems have been discussed. All the operations designated to obtaining new knowledge or data patterns in complex systems are algebraically described from the viewpoint of algebraic systems including predicates (in particular, finite ones) and operations on them. Thus, a system is considered to be complex if there is a great degree of uncertainty in the data and/or there are some serious problems with system’s decomposition.”

Keywords: big data, complex system, knowledge discovery, rough set, system decomposition, uncertainty in data

1 INTRODUCTION

Now there are no universal formal tools allowing us to describe and treat any kinds of complex systems. Nevertheless, we need to have some instruments and semantics to deal with systems complexity. Therefore, we try to consider the concept of a complex system from a logic-algebraic viewpoint, thereby considering it as a certain algebraic structure with predicates describing the available knowledge about the system and operations over them allowing us to describe some pieces of information in the system via the available knowledge structures. From our viewpoint, one of the most important features of a complex system is some level of uncertainty in data that does not allow making precise decisions based on the available information. In our opinion, the complexity lies not necessarily in a complicated database or knowledge base structure and not in a complicated and interconnected information flows. Also, the complexity does not necessarily mean many interconnected nodes in the system. We think that what does matter is the following: (a) the level of uncertainty we encounter when making decisions related to the system objects and processes; (b) the possibility of decomposing the system with losing, of course, some information on its elements. These two are associated with each other as the level of uncertainty obviously increases when we try to reduce the system to some subsystems and relations between them. Conversely, if system objects and processes become more vague and difficult to describe and deal with, the system’s decomposition to subsystems and relations between them becomes more difficult. These basic ideas have led us to an understanding that mathematical instruments dealing with vagueness and uncertainty in information systems should come into play. We have chosen rough set theory as the starting point for the following reasons. Other theories like fuzzy set theory or theory of evidence do allow treating uncertainty in data but do not allow convenient algebraic operations with information objects, which are necessary for understanding basic interrela-

ISSN: 1755-7437 (paper format), ISSN: 1755-7445 (online), http://www.witpress.com/journals
DOI: 10.2495/DNE-V13-N3-324-329

tions between them. Really, even the classical topological approach of Z. Pawlak extends the possibilities of probabilistic and statistical approaches. As we demonstrate in our papers, from our viewpoint, the opportunities of approximation based reasoning are far broader than Prof Pawlak supposed at the very beginning when creating his purely topological approach.

The founder of rough set theory Zdzislaw Pawlak considers his approach to modelling uncertainty and vagueness in data as very much promising for the future of intelligent industrial applications [1]. In his opinion confirmed by top quality research and practical results that have been obtained during the last decades, rough set models and tools are particularly suited for solving real-world problems in medicine, market and financial analysis, engineering, banking, pharmacology and other fields. Prof. Pawlak and his followers believe that rough set theory can be effectively used for modelling industrial processes and solving real-life problems in the following areas: material sciences, intelligent control, decision support systems, machine diagnosis, and neural networks.

From the viewpoint of prof. Pawlak, rough set methodology allows data mining and knowledge discovery by identifying relationships that cannot be found with the help of purely statistical approaches. Pawlak says that rough set based data reduction and assessing the importance of particular data seem to be very much useful in many branches of intelligent applications as well as rough set based generating logic rules that are easy to understand and interpret. We completely agree with this point of view.

Moreover, the approach outlined in the next section shows that the basic rough set ideology formulated by Prof Pawlak allows operating with Big Data, which are in principle heterogenous and multivariate. We do not argue with Prof Pawlak but follow his main ideas. These ideas have led us to some principally new approaches to treating uncertainty in data. What we suggest is this: (a) start with the available data structures and allowable operations on them; (b) describe any objects in a complex system via the well-known ones with the help of the available operations.

2 ALGEBRAIC APPROACH VERSUS KNOWLEDGE GRANULARITY BASED APPROACH

Almost everything that is related to rough set theory is based on so-called indiscernibility relations that induce some kinds of knowledge granularity, which allows classifying objects into similarity classes containing elements that cannot be discerned with respect to specific features. Of course, there is an obvious advantage of such a topological approach to treating uncertainty in data. As in the case of the classical database theory, the classical rough set theory by Z. Pawlak gives us a lot of opportunities to structure vagueness in our knowledge about the real world. In the case of rough sets, the indiscernibility relation and subsequent knowledge granularity allow bringing some order to data uncertainty, the possibility of operating with some basic “bricks” of the available information.

Nevertheless, the topological approach to treating vagueness in data implies some essential limitations as to our understanding of the uncertainty concept. When we encounter data points that cannot be discerned with respect to some features, we are uncertain about a class (cluster) to which different points should be attributed. This is how we understand the uncertainty concept following Pawlak’s approach.

In our series of papers [2–5] we try to give the rough set ideology a bit different perspective by avoiding the usage of the knowledge granularity concept. Whereas the classical rough set theory is based on the indiscernibility relation (which can be equivalence, tolerance or other types of relations), we do not use such a notion and try to consider uncertainty from a logic-algebraic viewpoint. Then a natural question arises: why do we think that we follow the rough set approach, if we do not accept the very first step in data analysis, namely the consideration of knowledge indiscernibility and granularity? The answer is this. We do not start with granularity. We start with rough approximations right away. We suppose that we originally have something (some knowledge about the world) expressed in terms of information objects, under which we understand not only data points but any predicates that can be defined on these elements (for example, on database records). It seems quite natural as any knowledge can be mathematically represented in the form of relations (predicates) or functions (that can be considered as a subtype of relations).

We introduce the concept of an approximation language, i.e. predicate operations allowed for describing dependencies between information objects. We describe information objects in terms of other objects as exactly as it is allowed by the approximation language. Thus the basic idea of Pawlak about approximation based reasoning stays intact. Moreover, it seems quite natural to start with approximating something that cannot be expressed exactly in terms of the available semantics. Following such an approach we do not need to suppose a priori which objects can be discerned with respect to some specific relations. All we need is just to find a logic-algebraic function that approximates the object under consideration “from below” and “from above”. An object under consideration can be any predicate, and according to our approach it should be expressed in terms of the other predicates available at the moment with the help of the predicate operations included in the approximation language.

The classical topological approach implies that there are the only upper and lower approximations for a rough set. We are not limited to a single approximation, but if an approximation cannot be improved in terms of the approximation language, we call it an exact approximation. According to what has been done before, we can now present a general definition of reasoning based on our rough set algebraic methodology, and we hope that industry professionals from various fields will be able to apply it to solve their specific knowledge discovery and machine learning problems outlined by Z.Pawlak and his followers.

1. Let $P_1(x_1,x_2,...,x_m),P_2(x_1,x_2,...,x_m),...,P_n(x_1,x_2,...,x_m)$ be predicates, i.e. functions taking on values 0 or 1. We interpret them as the accessible knowledge structure.
2. Let $Q_1(P_1,P_2,...,P_n),Q_2(P_1,P_2,...,P_n),...,Q_k(P_1,P_2,...,P_n)$ be operations over predicates, the set of which we call the approximation language. This language allows expressing information objects in terms of the accessible knowledge.
3. Let $X(x_1,x_2,...,x_m)$ be a predicate that we call an information object, which is required to be described in terms of the predicates $P_1,P_2,...,P_n$ with the help of the approximation language.
4. Any solution F of the functional equation $X(x_1,x_2,...,x_m)\rightarrow F(P_1,P_2,...,P_n)=1$, where F is constructed with the help of the approximation language, is called an upper approximation for X.
5. If a solution F^* of the above equation cannot be improved in the sense that any other solution F satisfies the equation $F^* \rightarrow F = 1$, then we call F^* the exact upper approximation for X.
6. Any solution F of the functional equation $F(P_1,P_2,...,P_n)\rightarrow X(x_1,x_2,...,x_m)=1$, where F is constructed with the help of the approximation language, is called a lower approximation for X.
7. If a solution F_* of the above equation cannot be improved in the sense that any other solution F satisfies the equation $F \rightarrow F_* = 1$, then we call F_* the exact lower approximation for X.
8. The function $F^* \wedge \overline{F_*}$ is called the boundary region for the object X.

The boundary region plays an important part in the classical Pawlak's rough set theory. It is a domain in which we are not certain to which set (class, cluster) an element should be attributed. If in the classical theory the boundary region is just a set of elements that we do not know exactly where to place, in our generalizations the boundary region is an object of a higher order. In the above model it is a function over predicates. This function along with the possibility of decomposing the system is suggested to be a measure of system's complexity. It should be noted that in case this function takes on only zero values, the object X under consideration can be precisely described in terms of the available objects with the help of the approximation language. If the boundary region represents a function equaling 1, it means that nothing can be said about the object under consideration in terms of the available knowledge. This is why we suggest this function as a measure (not the only one) of the "algebraic complexity" of a system.

It can be easily shown that the main properties of approximations in Pawlak's theory hold, but the above algebraic definitions allow us to generalize the classical Pawlak's concepts in different directions by variating approximation languages and the number of the original variables. For example, we can consider binary or ternary predicates and quantifiers as well as other operations on predicates.

The suggested approach allows explaining the meaning of an information object in terms of the accessible knowledge represented by predicates and allowable predicate operations. It is supposed that the predicates

$$P_1(x_1, x_2, ..., x_m), P_2(x_1, x_2, ..., x_m), ..., P_n(x_1, x_2, ..., x_m)$$

are obtained as a result of experiments and human intuition. As soon as these predicates and the necessary approximation language (semantics) are defined, the reasoning is carried out with the help of upper and lower approximations. Again, following the classical rough set theory approach, the exact upper approximation means a possible dependence, and the exact lower approximation means a necessary dependence. It should be noted that not only the exact approximations may be interesting for the investigation of hidden dependencies in data, but also other approximations may be useful. This point needs further research. We have shown that the exact approximations can be easily obtained in the case of unary predicates and Boolean operations [2], but in more complicated cases (binary predicates, quantifiers etc.) it may turn out that the exact approximations are difficult to calculate, and we can impose some limitations allowing us to select approximations close to the exact ones.

The boundary region in our understanding plays even a greater role in the sense that it allows extracting salient features from the available knowledge bases. The whole scheme of obtaining the importance of features is simple. We (a) eliminate a feature (a predicate in the general case) from consideration; (b) recalculate the upper and lower approximations *without* this predicate; (c) estimate how much the boundary region has changed as a result of this elimination. If the boundary region has changed very much (it can be measured), the eliminated feature is important. If the boundary region has changed a little bit, the feature is non-salient.

3 DECOMPOSITION OF A SYSTEM. MEASURE OF COMPLEXITY

As we mentioned in Introduction, not only the *uncertainty factor* but also difficulties related to decomposing the system are extremely important for defining the system as a complex one. In this connection we should note that describing a certain predicate in terms of some predicates depending on *sets of variables that do not intersect*, represents a degree of complexity associated with the *decomposition capacity factor*.

Consider a simple example. We should describe the predicate $X(x_1, x_2, ..., x_m)$ via some other predicates depending on the same variables. Suppose that we can find a predicate P "close" to X ("closeness" is a matter for a special discussion) such that

$P(x_1, x_2, ..., x_m) \rightarrow X(x_1, x_2, ..., x_m)$ or $X(x_1, x_2, ..., x_m) \rightarrow P(x_1, x_2, ..., x_m)$. If the predicate P can be decomposed to predicates of smaller dimensions (ideally $P = P_1(x_1)$ & $P_2(x_2)$ &….& $P_m(x_m)$, although the conjunction between these predicates is not the only option; we can consider any formula constructed with the help of the approximation language operations) so that the available knowledge is not lost but just decomposed, it means that the system can be simplified. If it is not possible, the system stays complex. The main idea is this. The harder the decomposition of the upper and lower approximations, the more complex the system is.

4 CONCLUSIONS

In this short paper we have outlined a mathematical (and, may be, philosophical) approach to understanding the complexity of an information system. We considered a general method of treating uncertainty in data, which is alternative to the classical rough set theory methodology, but it leads to approximation based reasoning with more general assumptions. As a rule, real-world intelligent applications have to deal with complicated data structures and Big Data. It means that a broader view on the rough set ideology can facilitate reasoning with elements of vagueness and uncertainty in the available knowledge. In our approach we do not require the existence of an indiscernibility relation in data, but we operate with the existing knowledge structure represented by sets of predicates and allowable predicate operations. We consider the complexity problem from two interrelated viewpoints: (a) how precisely we can describe an information object via the available and understandable objects; (b) how the obtained descriptions can be simplified by decomposing the available predicates. In our opinion the above two factors tell us a lot about the system complexity.

It should be noted that our approach cannot be considered as a universal one, although it is quite general. We have not considered here a lot of important issues: emergence, self-organization, adaptation etc. We have just concentrated on the following two aspects: uncertainty (this factor surely plays an important part in decision-making process in complex information systems); possibility of efficient decomposition for an information system objects and processes (in case of serious problems with such a decomposition the system should be recognized as a complex one). Even these matters have been considered in the general algebraic systems framework. Therefore, we hope that future research into these matters will be carried out and allow forming a strong general algebraic theory of complex systems.

REFERENCES

[1] Pawlak, Z., AI and intelligent industrial applications: the rough set perspective. *Cybernetics and Systems*, **31**, pp. 227–252, 2000. https://doi.org/10.1080/019697200124801

[2] Sitnikov, D. & Ryabov, O., An algebraic approach to defining rough set approximations and generating logic rules. *Data Mining V: Proc. 5th International Conference,* Malaga, Spain, 2004, pp. 179–188, 2004.

[3] Sitnikov, D., Ryabov, O., Titova, E. & Romanenko, O., A generalized algebraic approach to finding rough set approximations and generating logic rules. *Data Mining VIII: Proc. 8th International Conference,* Great Britain, pp. 3–12, 2007.

[4] Sitnikov, D., Titova, E., Ryabov, O. & Romanenko, O., An approach to finding reduced sets of information features describing discrete objects based on rough sets theory. *Data Mining IX: Proc. 9th International Conference,* Spain, pp. 3–11, 2008.

[5] Sitnikov, D., Titova, E. & Ryabov, O., A method for finding minimal sets of features adequately describing discrete information objects. *Data Mining X: Proc. 10th International Conference,* Greece, pp. 22–30, 2009.