A PRIMER ON ROUGH SETS: A NEW APPROACH TO DRAWING CONCLUSIONS FROM DATA

Zdzisław Pawlak∗

ABSTRACT

Rough set theory is a new mathematical approach to vague and uncertain data analysis. This Article explains basic concepts of the theory through a simple tutorial example and briefly outlines the application of the method to drawing conclusions from factual data. The presented approach can be used in some kind of legal reasoning.

INTRODUCTION

The problem of imperfect knowledge has been tackled for a long time by philosophers, logicians, and mathematicians. Recently, the problem also became a crucial issue for computer scientists, particularly in the area of artificial intelligence. There are many approaches to understanding and manipulating

∗ Zdzisław Pawlak obtained his Ph.D. in 1958 and D.Sc. in 1963 in the Theory of Computation from the Polish Academy of Sciences. He is a member of the Polish Academy of Sciences.

His research interests include cognitive sciences, decision support systems, inductive reasoning, vagueness, uncertainty, conflict analysis, logic, and philosophy of science. His previous interests were digital computers organization, information retrieval, and mathematical foundations of computer science. In the early 1950s he designed and supervised the construction of one of the first digital computers in Poland and Europe.

Professor Pawlak has also earned many awards, among them the State Award in Computer Science in 1978 and the Hugo Steinhaus award for achievements in applied mathematics in 1989. He is a member of several national and international societies and organizations, holds positions on the editorial boards of several dozen international journals, and is a program committee member of many national conferences on computer sciences.

Professor Pawlak has held over forty visiting university appointments in Europe, the United States, and Canada, and has participated in numerous conferences and seminars internationally. He has also published articles in international journals and several books on various aspects of computer science and application of mathematics. Professor Pawlak can be contacted at the Institute for Theoretical and Applied Informatics at the Polish Academy of Sciences.

imperfect knowledge. The most successful approach is, no doubt, Zadeh's fuzzy set theory.[1]

Rough set theory is another approach to this problem. From a philosophical point of view, rough set theory is a new approach to vagueness and uncertainty, and from a practical point of view, it is a new method of data analysis.[2]

The proposed method has the following important advantages:

- it provides efficient algorithms for finding hidden patterns in data;
- it finds reduced sets of data (data reduction);
- it evaluates significance of data;
- it generates minimal sets of decision rules from data;
- it is easy to understand;
- it offers straightforward interpretation of results;
- it can be used in both qualitative and quantitative data analysis; and
- it identifies relationships that would not be found using statistical methods.

Rough set theory overlaps with many other theories, such as fuzzy sets, evidence theory, and statistics. Nevertheless, it can be viewed in its own right as an independent, complementary, and noncompeting discipline.

The rough set methodology has found many real-life applications in various domains. It seems that the rough set approach can also be used in legal reasoning, particularly in drawing conclusions from factual data.

The rough set theory is based on sound mathematical foundations, but for simplicity's sake this Article refrains from advanced mathematical formalisms and tries to explain rudiments of the theory through a very simple example. The real life examples of applications are much more sophisticated and require more advanced extension of the theory.[3]

---

1 *See* Lotti Zadeh, *Fuzzy Sets*, 8 INFO. & CONTROL 338-53 (1965).

2 *See generally* ZDZISŁAW PAWLAK, ROUGH SETS: THEORETICAL ASPECTS OF REASONING ABOUT DATA (1991); Zdzisław Pawlak et al., *Rough Sets*, 38 COMM. ACM 88-95 (1995); Roman Slowinski, *Rough Set Approach to Decision Analysis*, 10 AI EXPERT 19 (1995).

3 For more information about rough sets and their applications, see TOSHINORI MUNAKATA, FUNDAMENTALS OF THE NEW ARTIFICIAL INTELLIGENCE: BEYOND TRADITIONAL PARADIGMS (1998); ROUGH FUZZY HYBRIDIZATION: A NEW TREND IN DECISION MAKING (S.K. Pal & A. Skowron eds., 1999); ROUGH SETS AND CURRENT TRENDS IN COMPUTING (L. Polkowski & A. Skowron eds., 1998); ROUGH SETS IN KNOWLEDGE DISCOVERY 1: METHODOLOGY AND APPLICATIONS (L. Polkowski & A.

I. AN EXAMPLE

Table 1 presents six facts concerning ninety-eight cases of driving a car in various driving conditions. In the table, columns labeled *weather, road*, and *time*—collectively called *condition attributes*—represent driving conditions. The column labeled *accident*, also called *decision attribute*, contains information regarding whether an accident has occurred in each case. *N* denotes the number of analogous cases.

Fact no.	*driving conditions*			*consequence*	*N*
	weather	*road*	*time*	*accident*	
1	*misty*	*icy*	*day*	*yes*	8
2	*foggy*	*icy*	*night*	*yes*	10
3	*misty*	*not icy*	*night*	*yes*	4
4	*sunny*	*icy*	*day*	*no*	50
5	*foggy*	*not icy*	*dusk*	*yes*	6
6	*misty*	*not icy*	*night*	*no*	20

Table 1

Table 1 illustrates the problem of finding the relationship between accidents and driving conditions, i.e., to describe the set of facts {1,2,3,5} (or the set of facts {4,6}) in terms of attributes *weather, road* and *time.* Note that the data are *inconsistent* because facts number 3 and 6 are inconsistent, i.e., they have the same conditions but different consequences, therefore the set of all accidents cannot be described in terms of attributes *weather, road* and *time.* However, we can describe the set of accidents approximately. To this end, an examination of the data reveals the following:

- {1,2,5} is the *maximal* set of facts that can *certainly* be classified as accidents in terms of the driving conditions;
- {1,2,3,5,6} is the set of all facts that *possibly* can be classified as accidents in terms of the driving conditions;
- {3,6} is the set of facts that can be classified neither as accident nor no accidents in terms of the driving conditions.

Skowron eds., 1998). Electronic Bulletin of the Rough Set Community, *at* http://www.cs.uregina.ca/~roughset (last visited Jan. 12, 2001); Grobian – The Rough Set Engine, *at* http://www.infj.ulst.ac.uk/~cccz23/grobian/grobian.html (last visited Jan. 12, 2001); The Rosetta Homepage, *at* http://www.idi.ntnu.no/~aleks/rosetta/ (last visited Jan. 12, 2001).

Note that the set {3,6} is the difference between sets {1,2,3,5,6} and {1,2,5}.

A. *Approximations*

The example provided above illustrates that some decisions cannot be described by means of conditions. However, they can be described with some approximations. Therefore, in what follows the following terminology is used:

- the set {1,2,5} is the *lower approximation* of the set {1,2,3,5};
- the set {1,2,3,5,6} is *the upper approximation* of the set {1,2,3,5};
- the set {3,6} is the *boundary region* of the set {1,2,3,5}.

Approximations are basic concepts of rough set theory and are used to draw conclusions from data.[4] Informal definitions of approximations are the following:

- the *lower approximation* of a set *X* with respect to data *D* is the set of all facts that can be *for certain* classified as *X* (are *certainly X*) in view of the data *D*;
- the *upper approximation* of a set *X* with respect to data *D* is the set of all facts that can be *possibly* classified as *X* (are *possibly X*) in view of the data *D*;
- the *boundary region* of a set *X* with respect to data *D* is the set of all facts that can be classified as neither *X* nor non-*X* in view of the data *D*.

Now we are able to say what rough sets are. A set *X* is *rough* (*approximate, inexact*) in view of the data *D* if its boundary region is nonempty; otherwise the set is *crisp* (*exact*).

Thus, the set of elements is rough (inexact) if it cannot be defined in terms of the data, i.e., it has some elements that can be classified neither as a member of the set nor its complement in view of the data.

B. *Data Reduction*

Another important issue in data analysis is reduction of data. Often, superfluous data can be removed from the data table while still allowing conclusions to be drawn from the data table. In order to reduce the data without affecting this property, we must preserve the consistency of the data. To this end we define the *degree of consistency* of a data table, which is given below:

[4] For precise, mathematical definitions of approximations, see sources cited *supra* note 3.

$$k = \frac{\text{the number of all consistent cases}}{\text{the number of all cases}}$$

Obviously, $0 \le k \le 1$.

A minimal subset of data that preserves consistency of the data is called a "reduct." For example, Tables 2 and 3 are reduced data tables obtained from Table 1.

Fact no.	*w*	*r*	*a*
1	*misty*	*icy*	*yes*
2	*foggy*	-	*yes*
3	*misty*	*not icy*	*yes*
4	*sunny*	-	*no*
5	*foggy*	-	*yes*
6	*misty*	*not icy*	*no*

Table 2

Fact no.	*w*	*t*	*a*
1	*Misty*	*day*	*yes*
2	*Foggy*	-	*yes*
3	*Misty*	*night*	*yes*
4	*Sunny*	-	*no*
5	*Foggy*	-	*yes*
6	*Misty*	*night*	*no*

Table 3

The algorithms for data reduction are rather sophisticated and this Article will not focus on this issue.[5]

C. *Decision Rules and Inverse Decision Rules*

In order to reason about data, we need a language of "decision rules," also known as "association rules" or "production rules." A decision rule is an implication in the form *if* Φ *then* Ψ, (in symbols $\Phi \rightarrow \Psi$), where Φ is called the "condition" and Ψ the "decision" of the rule. Φ and Ψ are logical formulas built up from attributes and attribute values and describe some properties of facts. Decision rules, on the other hand, express relationship between conditions and decisions.

Every fact in the data table determines a decision rule.

For example, Table 1 can be represented by the following set of decision rules:

(1) if (weather, misty) and (road, icy) and (time, day) then (accident, yes);

(2) if (weather, foggy) and (road, icy) and (time, night) then (accident, yes);

[5] For more about data reduction, see generally ROUGH SETS AND CURRENT TRENDS, *supra* note 3; ROUGH SETS IN KNOWLEDGE DISCOVERY, *supra* note 3.

(3) if (weather, misty) and (road, not icy) and (time, night) then (accident, yes);

(4) if (weather, sunny) and (road, icy) and (time, day) then (accident, no);

(5) if (weather, foggy) and (road, not icy) and (time, dusk) then (accident, yes);

(6) if (weather, misty) and (road, not icy) and (time, night) then (accident, no).

We can simplify the set of decision rules using Table 2 instead of Table 1:

(1′) if (weather, misty) and (road, icy) then (accident, yes)

(2′) if (weather, foggy) then (accident, yes)

(3′) if (weather, misty) and (road, not icy) then (accident, yes)

(4′) if (weather, sunny) then (accident, no)

(5′) if (weather, misty) and (road, not icy) then (accident, no)

We can get another set of decision rules employing Table 3.

Decision rules can be thought of as a formal language for drawing conclusions from data.

Sometimes we may be interested in *explanation* of decisions in terms of conditions. To this end, we need *inverse* decision rules which are obtained by mutually replacing conditions and decisions in every decision rule.

For example, the following inverse decision rules can be understood as an explanation of car accidents in terms of driving conditions:

(1″) if (accident, yes) then (road, icy) and (weather, misty);

(2″) if (accident, yes) then (weather, foggy);

(3″) if (accident, yes) then (road, not icy) and (weather, misty);

(4″) if (accident, no) then (weather, sunny);

(5″) if (accident, no) then (road, not icy) and (weather, misty).

Another explanation of accidents can be obtained by means of inverse decision rules obtained from Table 3. Thus, there is no unique explanation of accidents in view of the available data.

D. *Certainty and Coverage Factors*

Decision rules have interesting probabilistic properties that are discussed next.

With every decision rule $\Phi \rightarrow \Psi$, we associate two conditional probabilities:

- the certainty factor

$$\pi(\Psi \mid \Phi) = \frac{\text{number of all cases satisfying } \Phi \text{ and } \Psi}{\text{number of all cases satisfying } \Phi}$$

- the coverage factor

$$\pi(\Phi \mid \Psi) = \frac{\text{number of all cases satisfying } \Phi \text{ and } \Psi}{\text{number of all cases satisfying } \Psi}.$$

The certainty factor is the frequency of Ψs in Φ, and the coverage factor is the frequency of Φs in Ψ.

If a decision rule $\Phi \rightarrow \Psi$ uniquely determines decisions in terms of conditions, i.e., if $\pi(\Phi \mid \Psi) = 1$, then the rule is called "certain."

If a decision rule $\Phi \rightarrow \Psi$ does not determine decisions uniquely in terms of conditions, i.e., if $0 < \pi(\Phi \mid \Psi) < 1$, then the rule is called "uncertain."

For example:

"if (weather, misty) and (road, icy) and (time, day) then (accident, yes)" is a certain decision rule, whereas "if (weather, misty) and (road, not icy) and (time, night) then (accident, no)" is an uncertain decision rule.

Using Table 1, we can compute certainty and coverage factors for decision rules (1′) – (5′), which are presented in Table 4.

rule no.	*certainty*	*coverage*	*accident*
1′	1.00	0.29	*yes*
2′	1.00	0.57	*yes*
3′	0.17	0.14	*yes*
4′	1.00	0.71	*no*
5′	0.83	0.29	*no*

Table 4

Note that for inverse decision rules the certainty and coverage factors are mutually exhanged.

E. *Decision Rules and Approximations*

There is an interesting relationship between decision rules and approximations. *Certain* decision rules describe the *lower approximation* of the set of facts pointed out by the conclusion of the rule, whereas *uncertain* decision rules describe the *boundary region* of the set of facts pointed out by the conclusion of the rule.

For example, in Table 2 the following relationships between approximations and decision rules exist:

- Certain rules describing accidents (the lower approximation of the set of facts {1,2,3,5}):
  (1′) if (*weather*, *misty*) and (*road*, *icy*) then (*accident*, *yes*);
  (2′) if (*weather*, *foggy*) then (*accident, yes)*.
- Uncertain rule describing accidents (the boundary region {3,6}of the set of facts {1,2,3,5}):
  (3′) if (*weather*, *misty*) and (*road*, *not icy*) then (*accident*, *yes*).
- Certain rule describing lack of accidents (the lower approximation of the set of facts {4,6}):
  (4′) *if* (*weather*, *sunny*) *then* (*accident*, *no*)
- Uncertain rule describing lack of accidents (the boundary region {3,6}of the set of facts {4,6}):
  (5′) if (*weather, misty*) and (*road, not icy*) then (*accident, no*).

Another description of approximations can be obtained from Table 3. Because data reduction generally does not yield unique results, there is no unique description of approximations and boundary regions by means of decision rules.

F. *What the Data Tell Us*

From the decision rules (1′)–(5′) and the certainty factors, we can draw the following *conclusions:*

(1′) misty weather and icy road always caused accidents;
(2′) foggy weather always caused accidents;
(3′) misty weather and not icy road caused accidents in 17% of the cases;
(4′) sunny weather and icy road always caused safe driving;
(5′) misty weather and not icy road caused safe driving in 83% of the cases.

From the inverse decision rules (1″)-(5″) and the coverage factors we get the following *explanations*:

(1″) 29% of accidents occurred when the weather was misty and the road icy;
(2″) 57% of accidents occurred when the weather was foggy;

(3″) 14% of accidents occurred when the weather was misty and the road not icy;

(4″) 71% of safe driving took place when the weather was sunny;

(5″) 29% of safe driving took place when the weather was misty and the road not icy.

Summing up, from the decision rules (1′)-(5′) and the certainty factors lead to the following conclusions:

- misty weather and icy road or foggy weather *certainly* caused accidents;
- sunny weather and icy road *certainly* caused no accidents.
- misty weather and not icy road *most probably* caused no accidents.

The inverse decision rules (1″)-(5″) and the coverage factors led to the following explanations of driving accidents:

- the *most probable* cause of accidents is foggy weather;
- the *most probable* reason for the lack of accidents is sunny weather.

Other conclusions and explanations can be obtained by employing decision rules resulting from Table 3. Thus, as previously stated, there is no unique association of accidents with the driving conditions.

Note also that the data table represents a closed world, i.e., it is only a sample of a larger world. Therefore, the conclusions drawn are not universal but are valid only for the data. Whether they can be generalized depends on whether the data is a representative sample of a larger data set. But this Article does not discuss that problem, which is the central issue of inductive reasoning.

CONCLUSION

Rough set theory is a new method of drawing conclusions from data. The approach has found many nontrivial, real-life applications. It seems that rough set theory can be also used in some kinds of legal reasoning, but to this end more extensive research is necessary.