Int. J. Man–Machine Studies (1988) **29,** 81–95

Rough sets: probabilistic versus deterministic approach

ZDZISLAW PAWLAK†, S. K. M. WONG‡ AND WOJCIECH ZIARKO‡

† *Department of Complex Control Systems, Polish Academy of Sciences, Baltycka 5, 44–000 Gliwice, Poland, and ‡ Department of Computer Science, University of Regina, Regina, Saskatchewan, Canada, S4S 0A2*

(Received 14 September 1987)

1. Introduction

The issue of knowledge representation and the method of inferring decision rules are of fundamental nature in the design of intelligent systems. When knowledge of the system is *sufficient* and precise (without uncertainty), many problems in artificial intelligence can be successfully modelled by techniques such as first order logic (Kowalski, 1979; Barr & Feigenbanm, 1981). For instance, the logical structure of a computer program can be precisely described by a set of decision rules (Hurley, 1981). On the other hand, it is rather difficult, if possible at all, to describe unambiguously the knowledge and the decision-making process of a human expert such as a physician or a business manager. The knowledge acquired under these circumstances is often imprecise and incomplete. Many methods have been proposed to deal with the uncertain aspects inherent in a knowledge representation system. They vary from approaches based on subjective assignment to decision rules of some "certainty factors" (Shortliffe, 1976) to those based on fuzzy logic (Zadeh, 1981).

Recently, the notion of rough sets (Pawlak, 1982) was introduced, which provides a systematic framework for the study of the problems arising from imprecise and insufficient knowledge. Some of the advantages in using the rough-set concepts to expert systems design have been demonstrated by Pawlak, Slowinski & Slowinski (1986). However, in the existing rough-set model the probabilistic information crucial to non-deterministic classification (recognition) problems is not taken into consideration. For this reason, a probabilistic model has been proposed (Wong & Ziarko, 1986*a*, *b*) which is a natural extension of the rough-set method. The main advantage of the probabilistic model is that it provides a unified approach for both deterministic and non-deterministic knowledge representation systems. Furthermore, an effective inductive algorithm can be developed for a variety of applications (Wong & Ziarko, 1986*a*).

The main objective of this paper is to review and compare the fundamental results in the probabilistic and deterministic models of rough sets. In section 2, the basic ideas of rough sets are reviewed. We also present a method for simplifying a given knowledge representation system, which has been the subject of research for many years (Orlowska & Pawlak, 1984; Pawlak, 1984). In section 3, the probabilistic model is introduced and various concepts are explained. We conclude this section by

0020-7373/88/060081 + 15$03.00/0

highlighting the similarities between the probabilistic and deterministic models and demonstrate that all basic concepts of the deterministic theory have their equivalence in the probabilistic approach. In section 4 we focus on the differences between these two models in the context of their utility for inducing decision rules from a set of examples. Finally, we discuss some known and potential applications of both models in medical diagnosis, engineering control and machine learning systems.

2. Algebraic formalism of rough sets (deterministic case)

Before introducing the notions of the probabilistic rough-set model, we first review the basic concepts of rough sets proposed by Pawlak (1982).

2.1. BASIC CONCEPTS OF ROUGH SETS

Let U denote a finite set of objects, and let $R \subseteq U \times U$ be an equivalence relation on U. The pair $A = (U, R)$ is called an *approximation space*. If $(u, v) \in U$ and $u, v \in R$, we say that u and v are indistinguishable in A. R is referred to as an indiscernibility relation.

Let $R^* = \{X_1, X_2, \ldots, X_n\}$ denote the partition induced by the equivalence relation R, where X_i is an equivalence class of R (an elementary set of A). For any subset $X \subseteq U$, we can define the *lower* $\underline{A}(X)$ and *upper* $\bar{A}(X)$ approximations of X in the approximation space $A = (U, R)$ as follows:

$$\underline{A}(X) = \bigcup_{X_i \subseteq X} X_i$$

$$\bar{A}(X) = \bigcup_{X_i \cap X \neq \phi} X_i$$

That is, $\underline{A}(X)$ is the union of all those elementary sets in A, which are individually contained by X, whereas $\bar{A}(X)$ is the union of all those X_j of which has a non-empty intersection with X.

Given a subset $X \subseteq U$ representing certain *concept* of interest, we can characterize the approximation space $A = (U, R)$ with three distinct regions:

(1) $\underline{A}(X)$ is called the *positive* region $POS_A(X)$ of X in A;
(2) $\bar{A}(X) - \underline{A}(X)$ is called the *boundary* region $BND_A(X)$ of X in A;
(3) $U - \bar{A}(X)$ is called the *negative* region $NEG_A(X)$ of X in A.

Since, by assumption objects belonging to the same equivalence class of R are indistinguishable, it may be impossible to say with certainty if objects in the boundary region belong to X. In other words, the characterization of objects in X by the indiscernibility relation R is not precise enough if $BND_A(X) \neq \phi$. However, if $BND_A(X) = \phi$ (or $\underline{A}(X) = \bar{A}(X)$), we say that set X is *definable* in A; otherwise X is said to be a non-definable or a *rough* set. Note that any definable subset of U can be completely characterized by means of the elementary sets in A. In general, there are four different kinds of non-definable (rough) sets:

(1) set X is *roughly* definable if $\underline{A}(X) \neq \phi$ and $\bar{A}(X) \neq U$;
(2) set X is *internally* definable if $\underline{A}(X) \neq \phi$ and $\bar{A}(X) = U$;

(3) set X is *externally* definable if $\bar{A}(X) = \phi$ and $\bar{A}(X) \neq U$;
(4) set X is *totally* non-definable if $\bar{A}(X) = \phi$ and $\bar{A}(X) = U$.

Obviously, every (finite) union of elementary sets in an approximation space $A = (U, R)$ is definable. The collection $DEF(A)$ of all definable subsets in A uniquely defines a topological space $T_A = (U, DEF(A))$. The operation of the lower and upper approximations on a set X can, in fact, be interpreted respectively as the *interior* and *closure* operations in the topological space T_A.

From the topological interpretation of the approximation operations, for $X, Y \subseteq U$, one can easily obtain the following properties:

(1) $\underline{A}(\phi) = \bar{A}(\phi); \quad \underline{A}(U) = \bar{A}(U)$
(2) $\underline{A}(X) \subseteq X \subseteq \bar{A}(X)$
(3) $\underline{A}(X \cup Y) \supseteq \underline{A}(X) \cup \underline{A}(Y)$
(4) $\underline{A}(X \cap Y) = \underline{A}(X) \cap \underline{A}(Y)$
(5) $\bar{A}(X \cup Y) = \bar{A}(X) \cup \bar{A}(Y)$
(6) $\bar{A}(X \cap Y) \subseteq \bar{A}(X) \cap \bar{A}(Y)$
(7) $\underline{A}(U - X) = U - \bar{A}(X)$
(8) $\bar{A}(U - X) = U - \underline{A}(X)$
(9) $\underline{A}(\underline{A}(X)) = \bar{A}(\underline{A}(X)) = \underline{A}(X)$
(10) $\bar{A}(\bar{A}(X)) = \underline{A}(\bar{A}(X)) = \bar{A}(X)$.

2.2. ATTRIBUTE DEPENDENCY IN A KNOWLEDGE REPRESENTATION SYSTEM

The most natural application of rough sets is, perhaps, in those intelligent information systems in which the knowledge about a given set of objects can be characterized by the values of some selected attributes (features). Such a knowledge representation system (KRS) $S = (U, C, D, V, \rho)$ can be formally defined as follows:

U denotes a set of objects;
C is a set of condition attributes;
D is a set of action attributes;
$\rho : U \times F \to V$ is an information function, where $F = C \cup D$, $V = \bigcup_{a \in F} V_a$, and V_a is the domain of attribute $a \in F$.

Note that the restricted function, $\rho_u : F \to V$ defined by $\rho_u(a) = \rho(u, a)$ for every $u \in U$ and $a \in F$, provides the complete information about each object u in S.

An example of a knowledge representation system is given in Table 1. In this table the information we have about the objects (cars) in the universe $U = \{u_1, u_2, \ldots, u_8\}$ is characterized by means of the set $C = \{Size, Engine, Colour\}$ of condition attributes and the set $D = \{Max\text{-}Speed, Acceleration\}$ of decision attributes. The domains of the individual attributes are given by:

$$V_{Size} = \{compact, medium, full\},$$
$$V_{Engine} = \{diesel, gasoline, propane\},$$
$$V_{Colour} = \{black, white, silver\},$$
$$V_{Max\text{-}Speed} = \{low, medium, high\},$$
$$V_{Acceleration} = \{poor, good, excellent\},$$

TABLE 1
An example of a knowledge representation system

U	C			D	
Car	Size	Engine	Colour	Max-Speed	Acceleration
u_1	medium	diesel	silver	medium	poor
u_2	compact	gasoline	white	high	excellent
u_3	full	diesel	black	high	good
u_4	medium	gasoline	black	medium	excellent
u_5	medium	diesel	silver	low	good
u_6	full	propane	black	high	good
u_7	full	gasoline	white	high	excellent
u_8	compact	gasoline	white	low	good

The notion of attribute dependency plays an important role in pattern recognition, expert systems, and decision theory. Therefore, one of the main objectives in the analysis of a KRS is to investigate the dependency between the condition and action attributes.

For instance, in the context of the example presented in Table 1, we may be interested in knowing which attributes really determine the speed and acceleration of a car and which attributes are not important in this regard.

In a knowledge representation system $S = (U, C, D, V, \rho)$, for any subset G of condition attributes C or action attributes D, we can define an equivalence relation $\tilde{G}$ on U such that:

$$(u_i, u_j) \in \tilde{G} \text{ iff } \rho(u_i, g) = \rho(u_j, g) \qquad \text{for every} \quad g \in G.$$

Let $A \subseteq C$, $B \subseteq D$, and let $A^* = \{X_1, X_2, \ldots, X_n\}$ and $B^* = \{Y_1, Y_2, \ldots, Y_m\}$ denote the partitions on U induced respectively by the equivalence relations $\tilde{A}$ and $\tilde{B}$. An important question is that to what extent the partition B^* as a whole can be approximated or characterized by the partition A^*. Obviously, the quality of such an approximation depends very much on the relationship (dependency) between these two subsets of attributes A and B.

In terms of the lower and upper approximations $\underline{A}(Y_j)$ and $\bar{A}(Y_j)$ of $Y_j \in B^*$ in the approximation space $A = (U, \tilde{A})$, one can construct the positive, boundary, and negative regions of the partition B^* as follows:

$$POS_A(B^*) = \bigcup_{Y_j \in B^*} \underline{A}(Y_j),$$

$$BND_A(B^*) = \bigcup_{Y_j \in B^*} (\bar{A}(Y_j) - \underline{A}(Y_j)),$$

$$NEG_A(B^*) = U - \bigcup_{Y_j \in B^*} (\bar{A}(Y_j).$$

Based on the notions of rough sets, we can now define a plausible measure of the dependency of B on A by:

$$0 \leq \gamma_A(B) = |POS_A(B^*)| \,/\, |U| \leq 1$$

where $|\cdot|$ denotes the cardinality of a set. Note that $\gamma_A(B) = 1$ when B is *totally* dependent on A (i.e. A functionally determines B). If $0 < \gamma_A(B) < 1$, we say that B *roughly* depends on A. A and B are totally independent of each other when $\gamma_A(B) = 0$. In general, the dependency of B on A can be denoted by $A \overset{\gamma}{\rightarrow} B$.

For instance, it can be easily verified from Table 1 that:

$$\{Size, Engine, Colour\} \xrightarrow{0.5} Max\text{-}Speed,$$

$$\{Size, Engine\} \xrightarrow{0.5} Max\text{-}Speed,$$

$$\{Size\} \xrightarrow{0.375} Max\text{-}Speed.$$

This means that the knowledge represented by the values of the condition attributes *Size*, *Engine* and *Colour* is not sufficient to determine the speed of a car in all instances. It should also be noted that the attribute *Colour* is redundant or superfluous with respect to the attribute *Max-Speed* because the removal of the attribute *Colour* from the knowledge representation system would not affect the dependency between the set of the condition and decision attributes. We will discuss the problem of eliminating superfluous attributes in a knowledge representation system in subsection 2.4.

We wish to emphasize that the concept of attribute dependency defined above is an *algebraic* one. The notion of probabilistic dependency (independency) will be introduced in section 3.

2.3. DECOMPOSITION OF DECISION TABLES

Any knowledge representation system $S = (U, C, D, V, \rho)$ can be perceived as a *decision table* in which the values of attributes C stipulate the conditions for a particular decision as specified by the attribute values of D.

A decision table S can be classified according to the dependency measure $\gamma_C(D)$ as follows:

(i) S is deterministic if $\gamma_C(D) = 1$.
(ii) S is roughly deterministic if $0 < \gamma_C(D) < 1$.
(iii) S is totally non-deterministic if $\gamma_C(D) = 0$.

Clearly, in a totally non-deterministic decision table a number of possible actions may be taken for a given condition, while in a deterministic case each action is uniquely specified by a particular condition.

Any decision table can be decomposed *horizontally* into two sub-tables such that one is deterministic and the other is totally non-deterministic as shown in Example 1. (One of these sub-tables may, of course, be empty.)

2.3.1. Example 1

From the knowledge representation system given by Table 1 one obtains:

$$POS_C(D^*) = \{u_3, u_4, u_6, u_7\},$$
$$BND_C(D^*) = \{u_1, u_2, u_5, u_8\}.$$

TABLE 2
Deterministic decision table

U	C			D	
Car	Size	Engine	Colour	Max-Speed	Acceleration
u_3	full	diesel	black	high	good
u_4	medium	gasoline	black	medium	excellent
u_6	full	propane	black	high	good
u_7	full	gasoline	white	high	excellent

TABLE 3
Non-deterministic decision table

U	C			D	
Car	Size	Engine	Colour	Max-Speed	Acceleration
u_1	medium	diesel	silver	medium	poor
u_2	Compact	gasoline	white	high	excellent
u_5	medium	diesel	silver	low	good
u_8	Compact	gasoline	white	low	good

Since $C \xrightarrow{1} D$ holds in $POS_c(D^*)$ and $C \xrightarrow{0} D$ holds in $BND_C(D^*)$, Table 1 can be immediately decomposed into two sub-tables (Tables 2 & 3).

2.4. ELIMINATION OF SUPERFLUOUS ATTRIBUTES

In a knowledge representation system $S = (U, C, D, V, \rho)$ we describe each object by the attribute values of C. Very often some of the attributes in C may be redundant in the sense that they do not provide any additional information about the objects in S.

Let $B \subseteq C$ be a non-empty subset of condition attributes. We say that B is a *dependent* set of attributes if there exists a proper subset $B' \subset B$ such that $\bar{B}' = \bar{B}$, i.e. $B' \xrightarrow{1} B$; otherwise B is an *independent* set. B is said to be a *reduct* of C if B is a maximal independent set of condition attributes. In general, more than one reduct of C can be identified. The collection of all reducts of C will be denoted by $RED(C)$.

With the notion of positive regions, we can extend the above definitions to take into account the set of action attributes D. We say that $B \subseteq C$ is a dependent set *with respect to* D if there exists a proper subset $B' \subset B$ such that $POS_{B'}(D^*) = POS_B(D^*)$; otherwise B is regarded as an independent set with respect to D. A *relative* reduct B of C is defined to be a maximal independent set of condition attributes *with respect to* D. The collection of all such reducts will be denoted by $RED_D(C)$.

Note that for any reduct or relative reduct $\hat{C}$ of C, $C \xrightarrow{\gamma} D$ always implies $\hat{C} \xrightarrow{\gamma} D$. This observation provides us with an effective way to transform a decision table to a simpler one without any loss of information.

2.4.1. Example 2

It can be easily verified that $\hat{C} = \{Size, Engine\}$ is the only relative reduct of C in the KRS given by Table 1. It can therefore be transformed into another equivalent and simpler table (Table 4). It should be noted that there may exist more than one reduct of C. The set of attributes belonging to the intersection of all reducts of C:

$$CORE(C) = \bigcap_{B \in RED(C)} B$$

is referred to as the *core* of C.

An attribute $a \in C$ is said to be indispensable if $\tilde{C}_a \neq \tilde{C}$ for $C_a = C - \{a\}$. In fact, the core of C is equal to the union of all indispensable attributes in C.

Similarly, we can define the *relative core* of C with respect to the decision attributes D. An attribute $a \in C$ is said to be indispensable *with respect to* D if $POS_{C-\{a\}}(D^*) \neq POS_C(D^*)$. The *relative core* is equal to the intersection of all relative reducts, namely:

$$CORE_D(C) = \bigcap_{B \in RED_D(C)} D$$

which is the set of all indispensable attributes with respect to D.

The core can be easily computed from a KRS. Since every reduct contains the core, it is therefore advantageous to start with the core in order to find a reduct as illustrated below by the following example.

2.4.2. Example 3

The relative core of the set of attributes $C = \{Size, Engine, Colour\}$ with respect to the set $D' = \{Max\text{-}Speed\}$ is given by:

$$CORE_{D'}(C) = \{Size\}$$

There are two relative reducts of the set C with respect to D':

$$B_1 = \{Size, Engine\}$$
$$B_2 = \{Size, Colour\}$$

TABLE 4
Reduced decision table

N	C		D	
Car	Size	Engine	Max-Speed	Acceleration
u_1	medium	diesel	medium	poor
u_2	compact	gasoline	high	excellent
u_3	full	diesel	high	good
u_4	medium	gasoline	medium	excellent
u_5	medium	diesel	low	good
u_6	full	propane	high	good
u_7	full	gasoline	high	excellent
u_8	compact	gasoline	low	good

that is $RED_{D'}(C) = \{B_1, B_2\}$. Obviously,

$$CORE_{D'}(C) = B_1 \cap B_2$$

3. Probabilistic rough-set model

The algebraic concepts of rough sets as outlined in the previous section are not adequate to deal with information uncertainty inherent in many classification problems. This is primarily due to the fact that the rough-set model is based on a *deterministic* approach which deliberately ignores the available probabilistic information in its formalism.

In this section we attempt to formulate the notions of rough sets from a probabilistic point of view. Incorporating the probabilistic aspects into the algebraic rough-set model provides a flexible and useful framework for the study of non-deterministic systems.

3.1. PRELIMINARIES

Given a finite set of objects U, an equivalence relation R on U, and a probabilistic measure P defined on the σ-algebra of subsets of U, one can define a *probabilistic* approximation space a_P as a triple, $A_P = \langle U, R, P \rangle$. In this context, each subset of U corresponds to a random event representing a certain "concept" of interest.

Our primary objective here is to characterize an *expert* concept $Y \subseteq U$ in A_P by the known concepts $X_i (i = 1, 2, \ldots, n)$, the equivalence classes of R. Let $P(Y \mid X_i)$ denote the probabilty of occurrence of event Y conditioned on event X_i. In other words, $P(Y \mid X_i)$ is the probability that a randomly selected object with the description of concept X_i belongs to Y. In terms of these conditional probabilities, one can define $\underline{A}_P(Y)$ and $\bar{A}_P(Y)$, the lower and upper *probabilistic* approximations of Y in $a_P = \langle U, R, P \rangle$ as follows:

$$\underline{A}_P(Y) = \bigcup_{p(Y|X_i)>1/2} X_i$$

and

$$\bar{A}_P(Y) = \bigcup_{P(Y|X_i)\geq 1/2} X_i$$

Note that the above definitions are consistent with Bayes' decision procedure. Similarly to algebraic rough sets, we can partition the approximation space A_P into the *probabilisitc* positive, boundary, and negative regions of Y:

$$POS_{A_P}(Y) = \underline{A}_P(Y)$$

$$BND_{A_P}(Y) = \bar{A}_P(Y) - \underline{A}_P(Y) = \bigcup_{P(Y|X_i)=1/2} X_i$$

$$NEG_{A_P}(Y) = U - \bar{A}_P(Y) = \bigcup_{P(Y|X_i)<1/2} X_i$$

Whenever an object belongs to $POS_{A_P}(Y)$ (or $NEG_{A_P}(Y)$), one can conclude with some degree of confidence in a statistical sense that the object satisfies (or does not satisfy) concept Y. However, there is insufficient information for us to conclude whether an object in the boundary region matches concept Y or not.

If $\underline{A}_P(Y) = \bar{A}_P(Y)$ (i.e. $BND_{A_P}(Y) = \phi$), we say that the concept Y is *statistically* definable in the probabilistic approximation A_P; otherwise Y is statistically non-definable. Any definable set can be fully characterized by the elementary sets in A_P. A non-definable set is called a *statistically* rough set which can be classified into one of the following categories:

(1) Set Y is partially definable if $\underline{A}_P(Y) \neq \phi$ and $\bar{A}_P(Y) \neq U$.
(2) Set Y is internally definable if $\underline{A}_P(Y) \neq \phi$ and $\bar{A}_P(Y) = U$.
(3) Set Y is externally definable if $\underline{A}_P(Y) = \phi$ and $\bar{A}_P(Y) \neq U$.
(4) Set Y is totally non-definable if $\underline{A}_P(Y) = \phi$ and $\bar{A}_P(Y) = U$.

In contrast to the algebraic formalism presented in section 2, the family of statistically rough sets does not generate a topological space. This is due to the fact that for any two concepts X, $Y \subseteq U$, the identity $\bar{A}_P(X) \cup \bar{A}_P(Y) = \bar{A}_P(X \cup Y)$ does not necessarily hold in the probabilistic approximation space. Some of the useful properties are summarized below:

(1) $\underline{A}_P(\phi) = \bar{A}_P(\phi)$; $\quad \underline{A}_P(U) = \bar{A}_P(U)$
(2) $\underline{A}_P(X) \subseteq X \subseteq \bar{A}_P(X)$
(3) $\underline{A}_P(X \cup Y) \supseteq \underline{A}_P(X) \cup A_P(Y)$
(4) $\underline{A}_P(X \cap Y) \subseteq \underline{A}_P(X) \cap \underline{A}_P(Y)$
(5) $\bar{A}_P(X \cup Y) \supseteq \bar{A}_P(X) \cup \bar{A}_P(Y)$
(6) $\bar{A}_P(X \cap Y) \subseteq \bar{A}_P(X) \cap \bar{A}_P(Y)$

3.2. MEASURE OF STATISTICAL DEPENDENCY AND DEFINABILITY: THE INFORMATION DEPENDENCY

The idea of statistical dependency of random events defined here is fundamentally different from the standard notion of statistical dependency known in probability theory. In the standard approach the random events E_1 and E_2 are said to be independent if $P(E_2 \mid E_1) = P(E_2)$. This means that the probability of occurrence of the event E_2 has not been affected by the fact that event E_1 happened. In a similar way the independency of two random variables can be defined by considering all events corresponding to values of the first variable conditioned on the values of the second variable. If the requirement expressed by the above equation is not satisfied by some values of the random variables then the variables are said to be dependent and the degree of dependency can be determined using, for instance, the regression analysis. In the heart of such a definition of independency of random variables is the requirement of preservation of probability distribution. In other words, the conditional probability distribution of the first variable conditioned on each value of the second one is always the same as the unconditional distribution. However, this notion of independency (or the related notion of dependency) is useless when searching for such a notion of statistical dependency which would reasonably generalize the notion of functional (or partial functional) dependency. If the existence of functional dependency between variables (attributes) reflects our ability to determine values of one variable based on known values of the second then the adoption of the standard statistical dependency is totally incorrect because an example can very easily be constructed in which two variables are strongly dependent in functional sense and independent in statistical sense. Therefore, a new

dependency measure, referred to as an information dependency, has to be defined which combines statistical character with the ability to generalize the notion of functional dependency.

In this section we study the statistical information dependency versus the algebraic dependency. The notion of information dependency leads to a measure of definability (the approximate classification) of a concept in a probabilistic approximation space. More importantly, it also provides a simple interpretation of algebraic and statistical dependencies in a knowledge representation system.

Assume that there exists a probability measure P defined on the σ-algebra of subsets of U. We may regard any partition of U as a random variable. Let $X^* = \{X_1, X_2, \ldots, X_n\}$ and $Y^* = \{Y_1, Y_2, \ldots, Y_m\}$ denote the partitions induced, respectively, by two equivalence relations $\tilde{X}$ and $\tilde{Y}$ on U. We suggest here, according to information theory (Shannon, 1948), that the normalized entropy function $H(Y^* \mid X^*)$ defined by:

$$H(Y^* \mid X^*) = \sum_{i=1}^{n} P(X_i) H(Y^* \mid X_i)/\log m$$

where

$$H(Y^* \mid X_i) = -\sum_{j=1}^{m} P(Y_j \mid X_i) \log P(Y_j \mid X_i)$$

provides a plausible measure of information dependency of Y on X. The function $H(Y^* \mid X^*)$ satisfies the following important properties:

(1) $0 \leq H(Y^* \mid X^*) \leq 1$.
(2) Partition Y is *functionally dependent* on partition X if and only if $H(Y^* \mid X^*) = 0$.
(3) Partition Y is *completely independent* of partition X if and only if $H(Y^* \mid X^*) = 1$.

Thus, the conditional entropy function defined above provides a natural measure of the varying degree of definability of a concept or a set of concepts in a probabilistic approximation space. Given a concept $Y_1 \subseteq U$ in $A_p = \langle U, R, P \rangle$, we can define a partition $Y^* = \{Y_1, Y_2 = U - Y_1\}$ on U. Clearly, the entropy $H(Y^* \mid R^*)$ provides an overall measure of how well the concept Y_1 is being characterized by the partition $R^* = \{X_1, X_2, \ldots, X_n\}$ induced by R. It is interesting to note that $H(Y^* \mid R^*)$ is an upper bound of Bayes' classification error rate, namely:

$$\sum_{i=1}^{n} P(X_i) \max \{P(Y_1 \mid X_i), P(Y_2 \mid X_i)\} \leq \frac{1}{2} H(Y^* \mid R^*).$$

In particular, concept Y_1 is definable (algebraically) in $A = (U, R)$ if and only if $H(Y^* \mid R^*) = 0$, whereas Y_1 is totally non-definable (statistically) in $A_p = \langle U, R, P \rangle$ if and only if $H(Y^* \mid R^*) = 1$.

3.3. INFORMATION ATTRIBUTE DEPENDENCY (INDEPENDENCY)

The notion of information attribute dependency is in fact equivalent to that of partition dependency presented in the previous section. Let A^* and B^* be the

partitions induced by two arbitrary subsets of attributes in a knowledge representation system $S = (U, C, D, V, \rho)$. The degree of information dependency of set A on set B can be measured by the entropy $H(A^* \mid B^*)$:

$$B \xrightarrow{H(A^* \mid B^*)} A$$

3.3.1. Example 4

Consider the information dependency of the set of condition attributes $C = \{Size, Engine, Colour\}$ on $D = \{Max\text{-}Speed, Acceleration\}$ in the KRS given in Table 1. The dependency measure: $H(D^* \mid C^*) = 0{\cdot}215$ indicates strong (although not functional) relationship between C and D.

3.4. STATISTICAL NOTIONS OF REDUCT AND CORE

Similar to the algebraic formalism of rough sets, we can define, based on the idea of information dependency, the statistical notions of reduct and core as follows.

We say that a subset of condition attributes $A \subseteq C$ in S is a statistically dependent set if there exists a proper subset $B \subset A$ such that $H(A^* \mid B^*) = H(A^* \mid A^*)$; otherwise A is said to be a statistically independent set.

(1) *Reduct.*
A statistical reduct K of C is a maximal statistically independent subset of condition attributes. The collection of all statistical reducts of C will be denoted by $SRED(C)$. A subset of condition attributes $A \subseteq C$ is said to be a statistically dependent set in S *with respect to* D if there exists a proper subset $B \subset A$ such that $H(D^* \mid B^*) = H(D^* \mid A^*)$; otherwise A is a statistically independent set with respect to D.

(2) *Relative reduct.*
A statistical relative reduct K of C is a maximal statistically independent subset of condition attributes with respect to D. The collection of all such reducts will be denoted by $SRED_D(C)$. An attribute c is said to be statistically dispensable (superfluous) in C if $H(C^* \mid (C - \{c\})^*) = H(C^* \mid C^*)$; otherwise a is a statistically indispensable attribute in C.

Likewise, an attribute c is said to be statistically dispensable in c with respect to D if $H(D^* \mid (C - \{c\})^*) = H(D^* \mid C^*)$; otherwise c is a statistically indispensable attribute in C with respect to D.

(3) *Core.*
The statistical core is the set of all statistically indispensable condition attributes, which is in fact the intersection of all statistical reducts of C, namely:

$$SCORE(C) = \bigcap_{B \in SRED(C)} B$$

(4) *Relative core.*
The relative core is the set of statistically indispensable condition attributes with respect to D, which can also be written as the intersection of all relative reducts of C:

$$SCORE_D(C) = \bigcap_{B \in SRED(C)} B$$

Note that one can easily compute the reduct in a knowledge representation system. In many probabilistic classification problems, in order to simplify the decision rules, a reduct (preferably the smallest reduct consisting of a minimal number of attributes) need to be found. Since the set of attributes in the core appears in all reducts, it is therefore useful to start with the core in searching for a reduct.

3.4.1. Example 5

From the knowledge representation system given in Table 1, one obtains:

$$H(D^* \mid (C - \{Size\})^*) - H(D^* \mid C^*) = 0{\cdot}091$$
$$H(D^* \mid (C - \{Engine\})^*) - H(D^* \mid C^*) = 0{\cdot}0$$
$$H(D^* \mid (C - \{Colour\})^*) - H(D^* \mid C^*) = 0{\cdot}0$$

These results imply that either *Engine* or *Colour* is a superfluous attribute in C with respect to D. That is, removing attribute *Engine* or *Colour* from C does not affect the overall dependency of D on C. Consider the subset of attributes $C_1 = C = \{Colour\}$. Since:

$$H(D^* \mid (C_1 - \{Size\})^*) - H(D^* \mid C_1^*) = 0{\cdot}364$$
$$H(D^* \mid (C_1 - \{Engine\})^*) - H(D^* \mid C_1^*) = 0{\cdot}297$$

both *Size* and *Engine* are essential attributes in making decisions about the *Max-Speed* and *Acceleration* of a car. In fact, the statistical relative reduct and core are given by:

$$SRED_D(C) = \{Size, Engine\}$$
$$SCORE_D(C) = \{Size\}$$

4. Comparison of the probabilistic and algebraic rough-set models

In this section the main differences between the deterministic and probabilistic rough set models are discussed. In particular, we want to emphasize the limitations of the *deterministic* model, which have motivated the introduction of the probabilistic approach. In both models the common goal is to characterize a concept $Y \subseteq U$ in terms of the elementary concepts $X_1, X_2, \ldots, X_n$ in U. Depending on the application there are two possible scenarios:

(1) The universe of discourse U is known in the sense that we know the specifications of all objects in U;
(2) the universe of discourse U is known only partially, i.e. we know the specifications of objects in a subset $E \subset U$.

An application of type (1) is a decision table which contains all feasible combinations of conditions and associated actions. In this case, every object in the universe is known. However, the majority of applications of interest belong to type (2) in which we have only partial knowledge about the universe U. The problem is to find a characterization of some concept $Y \subseteq U$ based solely on the information contained in a test sample E. A typical application of this kind is machine learning where generalized decision rules are inferred from a training set of samples (Wong

& Ziarko 1986*a*). In this case, however, the data should be collected in a random fashion in order to ensure statistical validation of generated hypothesis.

The fundamental difference between the deterministic and probabilistic approaches lies in how the concept Y is characterized by the elementary concepts $X_1, X_2, \ldots, X_n$. The deterministic method limits itself to three-valued decision (*"yes"*, *"no"*, *"do not know"*) in the characterization of a concept. That is, a set of decision rules can be created to determine whether an object satisfying the specification $Des\,(X_i)$ of the known concept X_i also satisfies the specification $Des\,(Y)$ of the concept Y. In the deterministic case the decision rules can be written as

(1) $Des\,(X_i) \rightarrow Des\,(Y)$ if $X_i \subseteq POS\,(Y)$;
(2) $Des\,(X_i) \rightarrow not\ Des\,(Y)$ if $X_i \subseteq NEG\,(Y)$;
(3) $Des\,(X_i) \rightarrow$ *"unknown"* if $X_i \subseteq BND(Y)$.

4.1.1. Example 6

As an example, let us consider the collection of objects (cars) given in Table 1. Assume that we are interested in finding the characterization of the concept "good acceleration" in terms of the values of the attribute *Size*. The elementary concepts are "Size compact", "Size medium". It can be easily verified that the positive and negative regions of the concept "good acceleration" are both empty in the approximation space induced by values of the attribute *Size*. Thus, one obtains:

$$Size := medium \rightarrow \text{"}unknown\text{"};$$
$$Size := compact \rightarrow \text{"}unknown\text{"};$$
$$Size := full \rightarrow \text{"}unknown\text{"}.$$

It is clearly demonstrated by the above example that the deterministic rough-set method is not able to caputre and make use of the statistical information available in the boundary region (see Wong & Ziarko, 1986*b*, a detailed discussion of this problem). The statistical information is totally ignored by the deterministic method (which can, in fact, be justified in some applications, e.g. in the analysis of decision tables: Pawlak, 1986). In many other applications, such an approach is not adequate. Using the statistical information available in the boundary region, the probabilistic model is aimed at providing a more complete characterization of given concept Y.

In the probabilistic approach, the decision rules about a concept Y are given by:

(1) $Des\,(X_i) \xrightarrow{c} Des\,(Y)$ if $P(Y \mid X_i) > 0{\cdot}5$
(2) $Des\,(X_i) \xrightarrow{c} not\ Des\,(Y)$ if $P(Y \mid X_i) < 0{\cdot}5$
(3) $Des\,(X_i) \rightarrow$ *"unknown"* if $P(Y \mid X_i) = 0{\cdot}5$

where the certainty factor c for each rule is defined by $c = Max\,(P(Y \mid X_i), 1 - P(Y \mid X_i))$.

It can be easily seen that whenever a conclusive decision can be made using deterministic decision rules the same decision can be made with probabilistic decision rules. The converse is not true, however, as all probabilistic rules with a certainty factor $c < 1$ will be interpreted as *"unknown"* in the deterministic case.

4.1.2. Example 7

Let us compute probabilistic rules for the concept "good acceleration" in Example 6. From Table 1 we obtain the following rules:

$$Size := medium \xrightarrow{0.66} Acceleration := not\ good;$$

$$Size := Compact \longrightarrow \text{"}unknown\text{"};$$

$$Size := full \xrightarrow{0.66} Acceleration := good.$$

In contrast to the deterministic rules produced in the previous example, the above set of probabilistic rules can be used to predict the acceleration of a car based on its size.

5. Conclusions

A number of experimental systems have been implemented based on the deterministic rough-set theory. These applications include analysis of medical data of patients with duodenal ulcer (Pawlak, Slowinski & Slowinski, 1986), control algorithm acquisition in the process of cement kiln production (Mrozek, 1985), decision table analysis (Pawlak, 1986), pattern recognition (Wojcik, 1986) and approximate reasoning (Rasiowa & Skowron, 1986). The probabilistic model has proven to be a useful mathematical tool for dealing with some problems occurring in machine learning such as generation of decision rules from inconsistent training examples (Wong & Ziarko, 1986*a*) or training data analysis and reduction (the same applies to decision tables) (Ziarko, 1987). Experiments are under way on application of this model to isolated word recognition and for database design (Yasdi & Ziarko, 1987). The probabilistic model is also being used in experiments with design knowledge acquisition from artificially generated examples in the area of civil engineering (Arciszewski, Mustafa & Ziarko, 1987). The most comprehensive implementation of the probabilistic model is the system ANLYST (Ziarko, 1987) which is performing data analysis and reduction according to ideas presented in this article.

Most recent unpublished applications of the deterministic model involve a very successful system for airline pilot performance evaluation (this system has been adopted as standard by Polish airlines), geographical data classification with respect to terrain types, questionnaire analysis in sociology and psychology, and a variety of other medical applications. It must be stressed here that all these applications were implemented after standard statistical methods had been tried repeatedly and failed to provide satisfactory results. This, obviously, does not imply that rough set methods are better or can replace statistical methods. Instead the accumulated experience suggests that these two approaches are complementary to each other; in particular, the rough set methods are more justified and useful when the size of the set of experimental data is too small to apply standard statistical methods.

This research was supported in part by grants from the National Sciences and Engineering Research Council of Canada.

References

ARCISZEWSKI, T., MUSTAFA, M. & ZIARKO, W. (1987). A methodology of design knowledge acquisition for use in learning expert systems. *International Journal of Man–Machine Studies,* **27,** 23–32.

BARR, A. & FEIGENBANM, E. A. (1981). *The Handbook of Artificial Intelligence, Volume 1.* Willians Kaufman Inc.

HURLEY, R. B. (1981). *Decision Tables in Software Engineering.* Van Nostrand Reinhold.

KOWALSKI, R. (1979). *Logic for Problem Solving* Amsterdam: Elsevier.

MCDANIEL, H. (1978). *An Introduction to Decision Logic Tables.* New York: PBI (A Petrocelli Book).

MROZEK, A. (1985). Information systems and control algorithms. *Bulletin of the Polish Academy of Sciences,* **33.**

ORLOWSKI, E. & PAWLAK, Z. (1984). Expressive power of knowledge representation. *International Journal of Man–Machine Studies,* **20,** 485–500.

PAWLAK, Z. (1982). Rough sets. *International Journal of Information and Computer Sciences,* **11,** 145–172.

PAWLAK, Z. (1984). On superfluous attributes in knowledge representation. *Bulletin of the Polish Academy of Sciences,* **32.**

PAWLAK, Z. (1986). On decision tables. *Bulletin of the Polish Academy of Sciences,* **34,** 9–10.

PAWLAK, Z., SLOWINSKI, K. & SLOWINSKI, R. (1986). Rough classification of patients after highly selective vagotomy for duodendal ulcer. *International Journal of Man–Machine Studies,* **24,** 413–433.

RASIOWA, H. & SKOWRON, A. (1986). First step towards an approximation logic. *Journal of Symbolic Logic,* **51.**

SHANNON, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal,* **4.**

SHORTLIFFE, E. H. (1976). *Computer-Based Medical Consultations: MYCIN.* Amsterdam: Elsevier.

WOJCIK, Z. M. (1986). The rough sets utilization in linquistic pattern recognition. *Bulletin of the Polish Academy of Sciences,* **34.**

WONG, S. K. M. & ZIARKO, W. (1986*a*). INFER—an adaptive decision support system. *Proceedings of the 6th International Workshop On Expert Systems and Their Applications, Avignon, France.*

WONG, S. K. M. & ZIARKO, W. (1986*b*). Comparison of the probabilistic approximate classification and the fuzzy set model. *International Journal For Fuzzy Sets and Systems,* **21.**

YASDI, R. & ZIARKO, W. (1987). Conceptual schema design: a machine learning approach. *Proceedings of the 2nd ACM SIGART International Symposium on Methodologies for Intelligent Systems* (Charlotte 1987).

ZADEH, L. (1981). PRUF—a Meaning representional language for natural languages. In MAMDANI, E. & GAINES, B. *Fuzzy Reasoning And Its Applications.* London: Academic Press.

ZIARKO, W. (1987). On reduction of knowledge representation. *Proceedings of the 2nd ACM SIGART International Symposium on Methodologies for intelligent Systems* (Charlotte 1987) (Colloquia Series).