Article

Probabilistic approaches to rough sets

Y.Y. Yao

Department of Computer Science, University of Regina, Regina, Saskatchewan, Canada S4S 0A2
E-mail: yyao@cs.uregina.ca

Abstract: *Probabilistic approaches to rough sets in granulation, approximation and rule induction are reviewed. The Shannon entropy function is used to quantitatively characterize partitions of a universe. Both algebraic and probabilistic rough set approximations are studied. The probabilistic approximations are defined in a decision-theoretic framework. The problem of rule induction, a major application of rough set theory, is studied in probabilistic and information-theoretic terms. Two types of rules are analyzed: the local, low order rules, and the global, high order rules.*

Keywords: rough set approximations, granular computing, decision-theoretic rough set model, rule induction, high order rules, belief functions

1. Introduction

As a recently renewed research topic, granular computing is an umbrella term to cover any theories, methodologies, techniques and tools that make use of granules (i.e. subsets of a universe) in problem solving (Zadeh, 1979, 1997; Yao, 2000; Lin *et al.*, 2002). The basic guiding principle of granular computing is to 'exploit the tolerance for imprecision, uncertainty and partial truth to achieve tractability, robustness, low solution cost and better rapport with reality' (Zadeh, 1997). This principle offers a more practical philosophy for real-world problem solving. Instead of searching for the optimal solution, one may search for good approximate solutions.

The theory of rough sets provides a special and concrete model of granular computing (Pawlak, 1982, 1991). Three related issues of granulation, approximation and rule induction are central to studies of rough sets.

Granulation of a universe involves the decomposition of the universe into families of subsets, or the clustering of elements into groups. It leads to a collection of granules, with a granule being a clump of points (objects) drawn together by indistinguishability, similarity, proximity or functionality (Zadeh, 1997). Granulation may produce either a single-level flat structure or a multi-level hierarchical structure (Yao, 2001a). The theory of rough sets uses equivalence relations to represent relationships between elements of a universe. An equivalence relation induces a single-level granulation, namely, a partition of the universe.

A natural consequence of granulation is approximation. With respect to an equivalence relation, some subsets cannot be exactly expressed in terms of the equivalence classes and must be approximately represented by a pair of lower and upper approximations. By extending the approximations of subsets to a family of subsets, it is possible to study the approximation of a partition.

Rule induction deals with finding relationships between concepts. With each granule, or a family of granules, representing instances of a certain concept, one can study rule induction in set-theoretic terms (Yao, 2001b). Approximations of subsets and families of subsets offer insights and methods for rule induction (Pawlak, 1991).

Although mainstream research in rough set theory has been dominated by algebraic and non-probabilistic studies, probabilistic approaches have been applied to the theory ever since its inception (Wong & Ziarko, 1987; Pawlak *et al.*, 1988; Yao *et al.*, 1990; Yao & Wong, 1992; Düntsch & Gediga, 2001). More specifically, many authors implicitly used a probabilistic approach by counting the number of elements of a set. On the other hand, there is still a lack of systematic study of probabilistic approaches in a unified and general framework.

The main objective of this paper is to provide a critical analysis and review of probabilistic and information-theoretic approaches to rough sets. With respect to three related issues of granulation, approximation and rule induction, we focus the discussion on probability-related measures. Such a comprehensive study provides a solid basis for further study of probabilistic rough set theory.

2. Approximation space and information granulation

The underlying notion for granulation in rough sets is equivalence relations or partitions. Let U be a finite and nonempty universe. A binary relation $E \subseteq U \times U$ on U is called an equivalence relation if it is reflexive, symmetric and transitive. A partition of U is a collection of nonempty and pairwise disjoint subsets of U whose union is U. Each subset in a partition is also called a block. There is a one-to-

one relationship between equivalence relations and partitions. For an equivalence relation E, the equivalence class

$$[x]_E = \{y \in U | yEx\} \tag{1}$$

consists of all elements equivalent to x, and is the equivalence class containing the element x. The family of equivalence classes

$$U/E = \{[x]_E | x \in U\} \tag{2}$$

is a partition of the universe U. On the other hand, given a partition π of the universe, one can uniquely define an equivalence relation E_π:

$$xE_\pi y \Leftrightarrow x \text{ and } y \text{ are in the same block of the partition } \pi \tag{3}$$

In this paper, we will use equivalence relations and partitions interchangeably. The pair $\text{apr} = (U, E)$ is called an approximation space, indicating the intended application of the partition U/E for approximation (Pawlak, 1982).

The partition U/E is commonly known as the quotient set and provides a granulated view of the universe. A coarse-grained view of the universe may arise in several ways. For instance, the equivalence relation is derived based on available knowledge. Due to a lack of information or vague information, some distinct objects cannot be differentiated (Pawlak, 1982). That is, the available information only allows us to talk about an equivalence class as a whole instead of many individuals. In some situations, it may only be possible to observe or measure equivalence classes. It may also happen that a coarse-grained view is sufficient for a particular problem (Zhang & Zhang, 1992; Zadeh, 1997).

In the granulated view, equivalence classes are the basic building blocks and are called elementary (equivalence) granules. They are the smallest nonempty subsets that can be defined, observed or measured. From elementary granules, we can construct larger granules by taking unions of elementary granules. It is reasonable to assume that one can define, observe and measure these granules through the information and knowledge on the equivalence granules. The set of all definable granules, denoted by $\sigma(U/E)$, consists of the empty set $\varnothing$, the entire universe U, and unions of equivalence classes. The system $\sigma(U/E)$ is closed under set complement, intersection and union. It is a sub-Boolean algebra of the Boolean algebra formed by the power set 2^U of U and a σ-algebra of subsets of U generated by the family of equivalence classes U/E. In addition, U/E is the basis of the σ-algebra $\sigma(U/E)$.

Each partition represents one granulated view of the universe. Granulated views induced by all partitions form a partition lattice. The order relation of the lattice is defined as follows. A partition π_1 is a refinement of another partition π_2, or equivalently π_2 is a coarsening of π_1, denoted by $\pi_1 \preceq \pi_2$, if every block of π_1 is contained in some block of π_2, or equivalently each block of π_2 is a union of some blocks of π_1. In terms of equivalence relations, we have $U/E_1 \preceq U/E_2$ if and only if $E_1 \subseteq E_2$. Given two partitions π_1 and π_2, their meet $\pi_1 \wedge \pi_2$ is the largest partition which is a refinement of both π_1 and π_2, and their join $\pi_1 \vee \pi_2$ is the smallest partition which is a coarsening of both π_1 and π_2. The meet has all nonempty intersections of a block from π_1 and a block from π_2 as its blocks. The blocks of join are the smallest subsets which are exactly a union of blocks from both π_1 and π_2. In terms of equivalence relations, given two equivalence relations E_1 and E_2, the meet of U/E_1 and U/E_2 is defined by the equivalence relation $E_1 \cap E_2$, and the join is defined by the equivalence relation $(E_1 \cup E_2)^*$, the transitive closure of relation $E_1 \cup E_2$. The finest partition is given by $\{\{x\} | x \in U\}$ consisting of singleton subsets from U, and the coarsest partition is $\{U\}$.

The partition lattice clearly shows the structure of different granulations of the universe. It can be used to search for a suitable level of granulation for problem solving (Zhang & Zhang, 1992). Many machine learning algorithms using rough sets are based on the search of the partition lattice (Yao & Yao, 2002).

Information-theoretic measures can be used to quantify the degree of granularity of each partition (Lee, 1987; Düntsch & Gediga, 2001; Yao, 2003b). With respect to a partition $\pi = \{A_1, A_2, \ldots, A_m\}$, we have a probability distribution

$$P_\pi = \left(\frac{|A_1|}{|U|}, \frac{|A_2|}{|U|}, \ldots, \frac{|A_m|}{|U|}\right) \tag{4}$$

where $|\cdot|$ denotes the cardinality of a set. The Shannon entropy function of the probability distribution is defined by

$$H(\pi) = H(P_\pi) = -\sum_{i=1}^{m} \frac{|A_i|}{|U|} \log \frac{|A_i|}{|U|} \tag{5}$$

The entropy reaches the maximum value $\log |U|$ for the finest partition consisting of singleton subsets of U, and it reaches the minimum value 0 for the coarsest partition $\{U\}$. In general, for two partitions with $\pi_1 \preceq \pi_2$, we have $H(\pi_1) \geq H(\pi_2)$. That is, the value of the entropy correctly reflects the order of partitions with respect to their granularity.

Additional support for using the entropy as a measure of generality can be seen as follows. We can re-express equation (5) as

$$H(\pi) = \log |U| - \sum_{i=1}^{m} \frac{|A_i|}{|U|} \log |A_i| \tag{6}$$

The first term is a constant independent of any partition. The quantity $\log |A_i|$ is commonly known as the Hartley measure of information of the set A_i. It has been used to measure the amount of uncertainty associated with a finite set of possible alternatives, namely the nonspecificity

inherent in the set (Klir & Folger, 1988). The function log $|A_i|$ is a monotonic increasing transformation of the size of a set. It may be used to measure the granularity of the set. Large sets result in higher degrees of granularity than small sets. The second term of the equation is basically an expectation of granularity with respect to all subsets in a partition. It follows that we can use the following function as a measure of granularity for a partition:

$$G(\pi) = \sum_{i=1}^{m} \frac{|A_i|}{|U|} \log |A_i| \tag{7}$$

In contrast to the entropy function, for two partitions π_1 and π_2 with $\pi_1 \preceq \pi_2$, we have $G(\pi_1) \leq G(\pi_2)$. The coarsest partition $\{U\}$ has the maximum granularity value log $|U|$, and the finest partition $\{\{x\} | x \in U\}$ has the minimum granularity value 0.

3. Rough set approximations

In this section we discuss approximations of sets and approximations of probabilities, as well as probabilistic approximations of sets.

3.1. Approximations of sets

Consider an approximation space $\text{apr} = (U, E)$. The set of definable subsets is given by $\sigma(U/E)$. For a subset $A \subseteq U$, the greatest definable set contained in A is called the lower approximation of A, written $\underline{\text{apr}}_{U/E}(A)$, and the least definable set containing A is called the upper approximation of A, written $\overline{\text{apr}}_{U/E}(A)$. The subscript U/E indicates that the approximations are defined with respect to the partition U/E. When no confusion arises, we simply drop U/E. Lower and upper approximations can be expressed as

$$\begin{aligned} \underline{\text{apr}}(A) &= \bigcup \{X | X \in \sigma(U/E), X \subseteq A\} \\ \overline{\text{apr}}(A) &= \bigcap \{X | X \in \sigma(U/E), X \supseteq A\} \end{aligned} \tag{8}$$

In terms of equivalence classes, lower and upper approximations can be expressed by

$$\begin{aligned} \underline{\text{apr}}(A) &= \bigcup_{[x]_E \subseteq A} [x]_E \\ \overline{\text{apr}}(A) &= \bigcup_{[x]_E \cap A \neq \emptyset} [x]_E \end{aligned} \tag{9}$$

The lower approximation $\underline{\text{apr}}(A)$ is the union of equivalence classes which are subsets of A. The upper approximation $\overline{\text{apr}}(A)$ is the union of equivalence classes which have a nonempty intersection with A.

One may interpret $\underline{\text{apr}}, \overline{\text{apr}}: 2^U \to 2^U$ as two unary set-theoretic operators, called approximation operators. The system $(2^U, \neg, \underline{\text{apr}}, \overline{\text{apr}}, \cap, \cup)$ is called a Pawlak rough set algebra (Yao, 1996). It is an extension of the set algebra $(2^U, \neg, \cap, \cup)$. Properties of approximation operators, pertinent to our discussion, are summarized below:

(i) $\underline{\text{apr}}(A) = \neg\overline{\text{apr}}(\neg A)$,
$\overline{\text{apr}}(A) = \neg\underline{\text{apr}}(\neg A)$;

(ii) $\underline{\text{apr}}(A) = \overline{\text{apr}}(A) = A$, for $A \in \sigma(U/E)$;

(iii) $\underline{\text{apr}}(A) \subseteq A \subseteq \overline{\text{apr}}(A)$;

(iv) $\underline{\text{apr}}(A \cap B) = \underline{\text{apr}}(A) \cap \underline{\text{apr}}(B)$,
$\overline{\text{apr}}(A \cup B) = \overline{\text{apr}}(A) \cup \overline{\text{apr}}(B)$;

(v) $\underline{\text{apr}}(A \cup B) \supseteq \underline{\text{apr}}(A) \cup \underline{\text{apr}}(B)$,
$\overline{\text{apr}}(A \cap B) \subseteq \overline{\text{apr}}(A) \cap \overline{\text{apr}}(B)$.

Property (i) shows that lower and upper approximations are dual to each other. Property (ii) indicates that the lower and upper approximations of a definable set are the set itself. By property (iii), a set lies within its lower and upper approximations. Property (iv) states that the lower approximation distributes over intersection, and the upper approximation distributes over union. Property (v) shows the sub-distributivity of approximation operators.

Many probability-related measures on approximations have been proposed and studied. Pawlak (1982, 1991) suggested an accuracy measure of rough set approximation given by

$$\alpha(A) = \frac{|\text{apr}(A)|}{|\overline{\text{apr}}(A)|} = P(\underline{\text{apr}}(A) | \overline{\text{apr}}(A)) \tag{10}$$

It may be interpreted as the probability that an element belongs to the lower approximation, given that the element belongs to the upper approximation. This measure can also be expressed in terms of the well-known Marczewski–Steinhaus metric (Yao, 2001a). Measures of quality of lower and upper approximations are given respectively by Pawlak (1991):

$$\begin{aligned} \underline{q}(A) &= \frac{|\underline{\text{apr}}(A)|}{|U|} = P(\underline{\text{apr}}(A)) \\ \overline{q}(A) &= \frac{|\overline{\text{apr}}(A)|}{|U|} = P(\overline{\text{apr}}(A)) \end{aligned} \tag{11}$$

They are referred to as rough probability by Pawlak (1984) and have been used by many authors (Grzymala-Busse, 1987; Wong & Lingras, 1989; Yao & Lingras, 1998; Düntsch & Gediga, 2001). The relationship between the accuracy and quality of approximations can be expressed as

$$\alpha(A) = \frac{\underline{q}(A)}{\overline{q}(A)} \tag{12}$$

The accuracy measure can be re-expressed as

$$\alpha(A) = \frac{|\underline{\text{apr}}(A)|}{|\overline{\text{apr}}(A)|} = \frac{|\underline{\text{apr}}(A)|}{|U| - |\underline{\text{apr}}(\neg A)|} \tag{13}$$

which suggests that the accuracy measure also depends on the lower approximation of $\neg A$. Based on this observation, Gediga and Düntsch (2001) suggested use of the function

$$\underline{\gamma}(A) = \frac{|\underline{\text{apr}}(A)|}{|A|} = P(\underline{\text{apr}}(A)|A) \tag{14}$$

as a measure of the precision of deterministic approximation of A. Using the same argument, we suggest that the quality of non-deterministic approximation of A can be measured by

$$\overline{\gamma}(A) = \frac{|A|}{|\overline{\text{apr}}(A)|} = P(A|\overline{\text{apr}}(A)) \tag{15}$$

In this case, we have

$$\alpha(A) = \underline{\gamma}(A)\overline{\gamma}(A) \tag{16}$$

The two measures $\underline{q}$ and $\underline{\gamma}$ monotonically increase as $\underline{\text{apr}}(A)$ approaches A when different partitions are used. On the other hand, $\overline{q}$ and $\overline{\gamma}$ show the opposite direction of changes.

For two partitions π_1 and π_2 with $\pi_1 \preceq \pi_2$, we have

(vi) $\underline{\text{apr}}_{\pi_1}(A) \supseteq \underline{\text{apr}}_{\pi_2}(A)$, $\overline{\text{apr}}_{\pi_1}(A) \subseteq \overline{\text{apr}}_{\pi_2}(A)$.

By combining these with (iii), we obtain

$$\underline{\text{apr}}_{\pi_2}(A) \subseteq \underline{\text{apr}}_{\pi_1}(A) \subseteq A \subseteq \overline{\text{apr}}_{\pi_1}(A) \subseteq \overline{\text{apr}}_{\pi_2}(A) \tag{17}$$

As expected, a finer partition induces a tighter approximation. All of the measures correctly reflect this observation, as shown by the following properties:

(1) $\alpha_{\pi_1}(A) \geq \alpha_{\pi_2}(A)$;

(2) $\underline{q}_{\pi_1}(A) \geq \underline{q}_{\pi_2}(A)$, $\overline{q}_{\pi_1}(A) \leq \overline{q}_{\pi_2}(A)$;

(3) $\underline{\gamma}_{\pi_1}(A) \geq \underline{\gamma}_{\pi_2}(A)$, $\overline{\gamma}_{\pi_1}(A) \geq \overline{\gamma}_{\pi_2}(A)$.

That is, for two partitions with $\pi_1 \preceq \pi_2$, we obtain the same qualitative evaluation by all those measures, namely, π_1 is the same or better than π_2. For an arbitrary pair of partitions, the pair of $\underline{q}$ and $\underline{\gamma}$, or the pair of $\overline{q}$ and $\overline{\gamma}$, produce the same qualitative evaluation, which may be different from the one given by the α accuracy measure (Gediga & Düntsch, 2001).

The approximation of a subset can be easily extended to the approximation of a family of subsets. Consider two partitions $\pi_A = \{A_1, A_2, \ldots, A_n\}$ and $\pi_B = \{B_1, B_2, \ldots, B_m\}$ We construct an approximation space $\text{apr}_{\pi_A} = (U, E_{\pi_A})$ using the partition π_A. Each equivalence class of π_B is approximated by $\underline{\text{apr}}_{\pi_A}(B_i)$ and $\overline{\text{apr}}_{\pi_A}(B_i)$. By extending the accuracy and quality measure to the approximation of partition, Pawlak (1991) suggested the following quantities:

$$\begin{aligned} \alpha_{\pi_A}(\pi_B) &= \frac{\Sigma_{i=1}^m |\underline{\text{apr}}_{\pi_A}(B_i)|}{\Sigma_{i=1}^m |\overline{\text{apr}}_{\pi_A}(B_i)|} \\ \underline{\gamma}_{\pi_A}(\pi_B) &= \frac{\Sigma_{i=1}^m |\underline{\text{apr}}_{\pi_A}(B_i)|}{|U|} \end{aligned} \tag{18}$$

Furthermore, $\underline{\gamma}_{\pi_A}$ can be expressed as, respectively, the expectation, a weighted sum, and the sum of $\underline{\gamma}$, α and $\underline{q}$ on individual equivalence classes (Gediga & Düntsch, 2001):

$$\begin{aligned} \underline{\gamma}_{\pi_A}(\pi_B) &= \sum_{i=1}^m \frac{|B_i|}{|U|} \underline{\gamma}_{\pi_A}(B_i) = \sum_{i=1}^m P(B_i)\, \underline{\gamma}_{\pi_A}(B_i) \\ \underline{\gamma}_{\pi_A}(\pi_B) &= \sum_{i=1}^m \frac{|\overline{\text{apr}}_{\pi_A}(B_i)|}{|U|} \alpha_{\pi_A}(B_i) \\ &= \sum_{i=1}^m P(\overline{\text{apr}}_{\pi_A}(B_i)) \alpha_{\pi_A}(B_i) \\ \underline{\gamma}_{\pi_A}(\pi_B) &= \sum_{i=1}^m \underline{q}_{\pi_A}(B_i) \end{aligned} \tag{19}$$

The overall measure $\alpha_{\pi_A}(\pi_B)$ cannot be similarly expressed. It is reasonable to use as an alternative overall measure

$$\alpha'_{\pi_A}(\pi_B) = \sum_{i=1}^m \alpha_{\pi_A}(B_i) \tag{20}$$

In fact, $\alpha_{\pi_A}(\pi_B)$ and $\alpha'_{\pi_A}(\pi_B)$ represent two different averaging methods: one is the application of a measure to the pooled results, and the other is the average of the measurements on the individual results.

Given a subset $B \subseteq U$, we can partition the universe as $\{B, \neg B\}$. By applying the partition based measure $\underline{\gamma}_{\pi_A}$, Düntsch and Gediga (2001) suggested the following measure of the approximation quality of π_A with respect to B:

$$\underline{\gamma'}_{\pi_A}(B) = \underline{\gamma}_{\pi_A}(\pi_B) = \underline{q}_{\pi_A}(B) + \underline{q}_{\pi_A}(\neg B) \tag{21}$$

This measure is the ratio of correct classification of either B or $\neg B$ based on the partition π_A.

3.2. Approximations of probabilities

In an approximation space $\text{apr} = (U, E)$, suppose a set function is defined on $\sigma(U/E)$. One can extend the function to nondefinable subsets through the lower and upper approximations. Many authors have studied the approximation of probabilities in the framework of rough sets, which leads to belief functions (Pawlak, 1984; Grzymala-Busse, 1987; Skowron, 1989, 1990; Wong & Lingras, 1989; Skowron & Grzymala-Busse, 1994; Yao & Lingras, 1998).

A belief function is a mapping from 2^U to the unit interval [0, 1] and satisfies the following axioms.

(F1) $\text{Bel}(\varnothing) = 0$.

(F2) $\text{Bel}(U) = 1$.

(F3) For every positive integer n and every collection $A_1, \ldots, A_n \subseteq U$,

$$\begin{aligned} \text{Bel}(A_1 \cup A_2 \ldots \cup A_n) \geq & \sum_i \text{Bel}(A_i) \\ & - \sum_{i<j} \text{Bel}(A_i \cap A_j) \pm \ldots \\ & + (-1)^{n+1} \text{Bel}(A_1 \cap \ldots \cap A_n) \end{aligned}$$

Axioms (F1) and (F2) may be considered as normalization conditions. Axiom (F3) is a weaker version of the commonly known additivity axiom of probability functions. It is referred to as the axiom of superadditivity. The dual of a belief function, called a plausibility function Pl, is defined by

$$\mathrm{Pl}(A) = 1 - \mathrm{Bel}(\neg A) \tag{22}$$

For any subset $A \subseteq U$, $\mathrm{Bel}(A) \leq \mathrm{Pl}(A)$.

Consider first a simple case where a probability function P on 2^U is defined based on the counting of elements in a set (Grzymala-Busse, 1987; Skowron & Grzymala-Busse, 1994); namely, for $A \subseteq U$,

$$P(A) = \frac{|A|}{|U|} \tag{23}$$

Clearly, we have

$$\underline{q}(A) = P(\underline{\mathrm{apr}}(A)) \leq P(A) \leq P(\overline{\mathrm{apr}}(A)) = \overline{q}(A) \tag{24}$$

While $\underline{\mathrm{apr}}(A)$ and $\overline{\mathrm{apr}}(A)$ are approximations of the set A, $\underline{q}(A)$ and $\overline{q}(A)$ are the approximations of the probability of the set A. It can easily be verified by using properties (i)–(iv) that the qualities of lower and upper approximations are a pair of belief and plausibility functions.

Suppose P is a probability function defined on $\sigma(U/E)$. It is not defined for subsets of U which are not members of $\sigma(U/E)$. One can extend P to 2^U in two standard ways by defining functions P_* and P^*, traditionally called the inner measure and the outer measure induced by P. For an arbitrary subset $A \subseteq U$, we define

$$\begin{aligned} P_*(A) &= \sup\{P(X) | X \in \sigma(U/E), X \subseteq A\} = P(\underline{\mathrm{apr}}(A)) \\ P^*(A) &= \inf\{P(X) | X \in \sigma(U/E), X \supseteq A\} = P(\overline{\mathrm{apr}}(A)) \end{aligned} \tag{25}$$

Pawlak (1984) referred to the pair $(P_*(A), P^*(A))$ as the rough probability of A. The inner and outer probabilities P_* and P^* are a pair of belief and plausibility functions (Wong & Lingras, 1989; Fagin & Halpern, 1991; Yao & Lingras, 1998).

3.3. Probabilistic rough set approximations

Algebraic rough set approximations may be considered as qualitative approximations of a set. The extent of overlap between a set and an equivalence class is not considered. By incorporating the overlap, many authors have introduced and studied probabilistic rough set approximations (Wong & Ziarko, 1987; Pawlak *et al.*, 1988; Ziarko, 1993; Pawlak & Skowron, 1994). Most proposals introduced certain parameters based on intuitive arguments. The decision-theoretic rough set model provides a solid basis for probabilistic approximations (Yao *et al.*, 1990; Yao & Wong, 1992).

We first briefly review the Bayesian decision-theoretic framework (Duda & Hart, 1973). Let $\Omega = \{\omega_1, \ldots, \omega_s\}$ be a finite set of s states, and let $\mathscr{A} = \{a_1, \ldots, a_m\}$ be a finite set of m possible actions. Let $P(\omega_j|\boldsymbol{x})$ be the conditional probability of an object x being in state ω_j given that the object is described by $\boldsymbol{x}$. Let $\lambda(a_i|\omega_j)$ denote the loss, or cost, for taking action a_i when the state is ω_j. For an object with description $\boldsymbol{x}$, suppose an action a_i is taken. Since $P(\omega_j|\boldsymbol{x})$ is the probability that the true state is ω_j given $\boldsymbol{x}$, the expected loss associated with taking action a_i is given by

$$R(a_i|\boldsymbol{x}) = \sum_{j=1}^{s} \lambda(a_i|\omega_j) P(\omega_j|\boldsymbol{x}) \tag{26}$$

The quantity $R(a_i|\boldsymbol{x})$ is called the conditional risk. Given a description $\boldsymbol{x}$, a decision procedure is a function $\tau(\boldsymbol{x})$ that specifies which action to take. For every $\boldsymbol{x}$, $\tau(\boldsymbol{x})$ chooses one action from $a_1, \ldots, a_m$. The overall risk $\boldsymbol{R}$ is the expected loss associated with a given decision procedure. Since $R(\tau(\boldsymbol{x})|\boldsymbol{x})$ is the conditional risk associated with action $\tau(\boldsymbol{x})$, the overall risk is defined by

$$\boldsymbol{R} = \sum_{\boldsymbol{x}} R(\tau(\boldsymbol{x})|\boldsymbol{x}) P(\boldsymbol{x}) \tag{27}$$

where the summation is over the set of all possible descriptions of objects.

One can obtain an optimal decision procedure by minimizing the overall risk. If $\tau(\boldsymbol{x})$ is chosen so that $R(\tau(\boldsymbol{x})|\boldsymbol{x})$ is as small as possible for every $\boldsymbol{x}$, the overall risk $\boldsymbol{R}$ is minimized. The Bayesian decision procedure can therefore be formally stated as follows. For every $\boldsymbol{x}$, compute the conditional risk $R(a_i|\boldsymbol{x})$ for $i = 1, \ldots, m$ defined by equation (26), and then select the action for which the conditional risk is the minimum. If more than one action minimizes $R(a_i|\boldsymbol{x})$, any tie-breaking rule can be used.

The Bayesian decision procedure can be applied to define probabilistic rough set approximations. Given a subset $A \subseteq U$, we can form a set of two states $\Omega = \{A, \neg A\}$ indicating that an element is in A and not in A, respectively. We use the same symbol to denote both a subset A and the corresponding state. In the non-probabilistic rough set model, with respect to A, we divide the universe U into three disjoint regions, the positive region POS(A), the negative region NEG(A) and the boundary region BND(A):

$$\begin{aligned} \mathrm{POS}(A) &= \underline{\mathrm{apr}}(A) \\ \mathrm{NEG}(A) &= U - \overline{\mathrm{apr}}(A) \\ \mathrm{BND}(A) &= \overline{\mathrm{apr}}(A) - \underline{\mathrm{apr}}(A) \end{aligned} \tag{28}$$

In developing a probabilistic rough set model, with respect to three regions, the set of actions is given by $\mathscr{A} = \{a_1, a_2, a_3\}$, where a_1, a_2 and a_3 represent the three actions in classifying an object, deciding POS(A), deciding NEG(A) and deciding BND(A), respectively. The symbol $[x]_E$, the equivalence class containing x, is also used to represent a description of x. The required conditional probabilities are defined by the rough membership functions (Pawlak & Skowron, 1994)

$$\mu_A(x) = \frac{|[x]_E \cap A|}{|[x]_E|} = P(A|[x]_E) \tag{29}$$

Let $\lambda(a_i|A)$ denote the loss incurred for taking action a_i when an object in fact belongs to A, and let $\lambda(a_i|\neg A)$ denote the loss incurred for taking the same action when the object does not belong to A. The expected loss $R(a_i|[x]_E)$ associated with taking the individual actions can be expressed as

$$\begin{aligned} R(a_1|[x]_E) &= \lambda_{11}P(A|[x]_E) + \lambda_{12}P(\neg A|[x]_E) \\ R(a_2|[x]_E) &= \lambda_{21}P(A|[x]_E) + \lambda_{22}P(\neg A|[x]_E) \\ R(a_3|[x]_E) &= \lambda_{31}P(A|[x]_E) + \lambda_{32}P(\neg A|[x]_E) \end{aligned} \tag{30}$$

where $\lambda_{i1} = \lambda(a_i|A)$, $\lambda_{i2} = \lambda(a_i|\neg A)$ and $i = 1, 2, 3$. The Bayesian decision procedure leads to the following minimum-risk decision rules.

(P) If $R(a_1|[x]_E) \leq R(a_2|[x]_E)$ and $R(a_1|[x]_E) \leq R(a_3|[x]_E)$, decide POS($A$).
(N) If $R(a_2|[x]_E) \leq R(a_1|[x]_E)$ and $R(a_2|[x]_E) \leq R(a_3|[x]_E)$, decide NEG($A$).
(B) If $R(a_3|[x]_E) \leq R(a_1|[x]_E)$ and $R(a_3|[x]_E) \leq R(a_2|[x]_E)$, decide BND($A$).

Tie-breaking rules should be added so that each element is classified into only one region. Since $P(A|[x]_E) + P(\neg A|[x]_E) = 1$, the above decision rules can be simplified such that only the probabilities $P(A|[x]_E)$ are involved. We can classify any object in the equivalence class $[x]_E$ based only on the probabilities $P(A|[x]_E)$, i.e. the rough membership values, and the given loss function λ_{ij} ($i = 1, 2, 3$; $j = 1, 2$).

Consider a special kind of loss function with $\lambda_{11} \leq \lambda_{31} < \lambda_{21}$ and $\lambda_{22} \leq \lambda_{32} < \lambda_{12}$. That is, the loss of classifying an object x belonging to A into the positive region POS(A) is less than or equal to the loss of classifying x into the boundary region BND(A), and both of these losses are strictly less than the loss of classifying x into the negative region NEG(A). The reverse order of losses is used for classifying an object that does not belong to A. For this type of loss function, the minimum-risk decision rules (P), (N), (B) can be written as

(P) if $P(A|[x]_E) \geq \gamma$ and $P(A|[x]_E) \geq \alpha$, decide POS(A);
(N) if $P(A|[x]_E) \leq \beta$ and $P(A|[x]_E) \leq \gamma$, decide NEG(A);
(B) if $\beta \leq P(A|[x]_E) \leq \alpha$, decide BND($A$);

where

$$\begin{aligned} \alpha &= \frac{\lambda_{12} - \lambda_{32}}{(\lambda_{31} - \lambda_{32}) - (\lambda_{11} - \lambda_{12})} \\ \gamma &= \frac{\lambda_{12} - \lambda_{22}}{(\lambda_{21} - \lambda_{22}) - (\lambda_{11} - \lambda_{12})} \\ \beta &= \frac{\lambda_{32} - \lambda_{22}}{(\lambda_{21} - \lambda_{22}) - (\lambda_{31} - \lambda_{32})} \end{aligned} \tag{31}$$

By the assumptions $\lambda_{11} \leq \lambda_{31} < \lambda_{21}$ and $\lambda_{22} \leq \lambda_{32} < \lambda_{12}$, it follows that $\alpha \in (0, 1]$, $\gamma \in (0, 1)$ and $\beta \in [0, 1)$.

A loss function should be chosen in such a way as to satisfy the condition $\alpha \geq \beta$. This ensures that the results are consistent with rough set approximations. Namely, the lower approximation is a subset of the upper approximation, and the boundary region may be nonempty. When $\alpha > \beta$, we have $\alpha > \gamma > \beta$. After tie-breaking, we obtain the decision rules

(P1) if $P(A|[x]_E) \geq \alpha$, decide POS(A);
(N1) if $P(A|[x]_E) \leq \beta$, decide NEG(A);
(B1) if $\beta < P(A|[x]_E) < \alpha$, decide BND($A$).

When $\alpha = \beta$, we have $\alpha = \gamma = \beta$. In this case, we use the decision rules

(P2) if $P(A|[x]_E) > \alpha$, decide POS(A);
(N2) if $P(A|[x]_E) < \alpha$, decide NEG(A);
(B2) if $P(A|[x]_E) = \alpha$, decide BND(A).

For the second set of decision rules, we use a tie-breaking criterion so that the boundary region may be nonempty.

The standard and other probabilistic rough set models can be easily derived by choosing different loss functions. Consider the loss function

$$\lambda_{12} = \lambda_{21} = 1 \qquad \lambda_{11} = \lambda_{22} = \lambda_{31} = \lambda_{32} = 0 \tag{32}$$

There is a unit cost if an object belonging to A is classified into the negative region or if an object not belonging to A is classified into the positive region; otherwise there is no cost. In this case, we have $\alpha = 1 > \beta = 0$, $\alpha = 1 - \beta$ and $\gamma = 0.5$. According to decision rules (P1), (N1), (B1), we obtain the standard rough set approximations (Pawlak, 1982, 1991). Another loss function is given by

$$\lambda_{12} = \lambda_{21} = 1 \qquad \lambda_{31} = \lambda_{32} = 0.5 \qquad \lambda_{11} = \lambda_{22} = 0 \tag{33}$$

A unit cost is incurred if the system classifies an object belonging to A into the negative region or an object not belonging to A into the positive region; half of a unit cost is incurred if any object is classified into the boundary region. There is no cost for other cases. It follows that $\alpha = \beta = \gamma = 0.5$. By using decision rules (P2), (N2), (B2), we obtain the probabilistic rough set approximation proposed by Pawlak *et al.* (1988).

Suppose a loss function with $\lambda_{11} \leq \lambda_{31} < \lambda_{21}$ and $\lambda_{22} \leq \lambda_{32} < \lambda_{12}$ satisfies the conditions

$$\begin{aligned} \lambda_{12} - \lambda_{32} &\geq \lambda_{31} - \lambda_{11} \\ (\lambda_{12} - \lambda_{32})(\lambda_{32} - \lambda_{22}) &= (\lambda_{31} - \lambda_{11})(\lambda_{21} - \lambda_{31}) \end{aligned} \tag{34}$$

We have $\alpha = 1 - \beta \geq 0.5$. This leads to the variable precision rough set model (Ziarko, 1993).

4. Probabilistic measures for rule induction

An important application of rough sets is data analysis and rule induction (Wong & Ziarko, 1986; Pawlak, 1991; Grzymala-Busse, 1992; Tsumoto, 1998). This section reviews probabilistic and information-theoretic measures used in rule induction algorithms (Yao & Zhong, 1999; Yao, 2003b).

4.1. Information tables

An information table provides a convenient way to describe a finite set of objects by a finite set of attributes (Pawlak, 1991). In this paper, we use an extended information table by adding binary relations on attribute values, and two languages (Yao, 2001b). Formally, an information table can be expressed as

$$S = (U, \mathrm{At}, L_v, L_r, \{V_a | a \in \mathrm{At}\}, \{R_a | a \in \mathrm{At}\}, \{I_a | a \in \mathrm{At}\})$$

where U is a finite nonempty set of objects, At is a finite nonempty set of attributes, L_v is a language dealing with values of objects, L_r is a language dealing with relations of objects, V_a is a nonempty set of values for $a \in \mathrm{At}$, R_a is a nonempty set of binary relations on V_a, $a \in \mathrm{At}$, and I_a: $U \rightarrow V_a$ is an information function. Each information function I_a is a total function that maps an object of U to exactly one value in V_a. An information table represents all available information and knowledge. That is, objects are only perceived, observed or measured by using a finite number of properties.

In the language L_v, an atomic formula is given by $(a, \mathscr{R}, v)$, where $a \in \mathrm{At}$, $\mathscr{R} \in R_a$ and $v \in V_a$. If ϕ and ψ are formulae, then so are $\neg\phi$, $\phi \wedge \psi$, $\phi \vee \psi$, $\phi \rightarrow \psi$ and $\phi \equiv \psi$. The semantics of the language L_v can be defined in the Tarski style through the notions of a model and satisfiability. The model is an information table S, which provides an interpretation for symbols and formulae of L_v. The satisfiability of a formula ϕ by an object x, written $x \models_s \phi$ or in short $x \models \phi$ if S is understood, is given by the following conditions:

(m1) $x \models (a, \mathscr{R}, v)$ iff $I_a(x)\ \mathscr{R} v$,
(m2) $x \models \neg\phi$ iff not $x \models \phi$,
(m3) $x \models \phi \wedge \psi$ iff $x \models \phi$ and $x \models \psi$,
(m4) $x \models \phi \vee \psi$ iff $x \models \phi$ or $x \models \psi$,
(m5) $x \models \phi \rightarrow \psi$ iff $x \models \neg\phi \vee \psi$,
(m6) $x \models \phi \equiv \psi$ iff $x \models \phi \rightarrow \psi$ and $x \models \psi \rightarrow \phi$.

If ϕ is a formula, the set $m_s(\phi)$ defined by

$$m_S(\phi) = \{x \in U | x \models \phi\} \tag{35}$$

is called the meaning of the formula ϕ in S. If S is understood, we simply write $m(\phi)$. The following properties hold:

(a) $m(a, \mathscr{R}, v) = \{x \in U | I_a(x)\ \mathscr{R}\ v\}$,
(b) $m(\neg\phi) = \neg m(\phi)$,
(c) $m(\phi \wedge \psi) = m(\phi) \cap m(\psi)$,
(d) $m(\phi \vee \psi) = m(\phi) \cup m(\psi)$,
(e) $m(\phi \rightarrow \psi) = \neg m(\phi) \cup m(\psi)$,
(f) $m(\phi \equiv \psi) = (m(\phi) \cap m(\psi)) \cup (\neg m(\phi) \cap \neg m(\psi))$.

The meaning of a formula ϕ is therefore the set of all objects having the property expressed by the formula ϕ. In other words, ϕ can be viewed as the description of the set of objects $m(\phi)$. Thus, a connection between formulae of L_v and subsets of U is established. When the relation $\mathscr{R}$ is chosen to be the equality relation $=$, we obtain the conventional decision logic language (Pawlak, 1991).

With the introduction of language L_v, we have a formal description of concepts. A concept definable in an information table is a pair $(\phi, m(\phi))$, where $\phi \in L_v$. More specifically, ϕ is a description of $m(\phi)$ in S, the intension of concept $(\phi, m(\phi))$, and $m(\phi)$ is the set of objects satisfying ϕ, the extension of concept $(\phi, m(\phi))$. A concept $(\phi, m(\phi))$ is said to be a sub-concept of another concept $(\psi, m(\psi))$, or $(\psi, m(\psi))$ a super-concept of $(\phi, m(\phi))$, if $m(\phi) \subseteq m(\psi)$. A concept $(\phi, m(\phi))$ is said to be the smallest nonempty concept in S if there does not exist another proper nonempty sub-concept of $(\phi, m(\phi))$. Two concepts $(\phi, m(\phi))$ and $(\psi, m(\psi))$ are disjoint if $m(\phi) \cap m(\psi) = \emptyset$. If $m(\phi) \cap m(\psi) \neq \emptyset$, we say that the two concepts have a nonempty overlap and hence are related.

The language L_r is defined in a similar manner to L_v, except that an atomic formula is given by $(a, \mathscr{R})$, where $\mathscr{R} \in R_a$ and $a \in \mathrm{At}$. Semantics of formulae of L_r are interpreted by pairs of objects in U. That is,

(m1′) $(x, y) \models (a, \mathscr{R})$ iff $I_a(x)\ \mathscr{R}\ I_a(y)$.

For formula ϕ, the set $m_s(\phi)$ defined by

$$m_S(\phi) = \{(x, y) \in U \times U | (x, y) \models \phi\} \tag{36}$$

is called the meaning set of ϕ in S. If S is understood, we simply write $m(\phi)$. A pair $(x, y) \in m(\phi)$ is said to satisfy the expression ϕ. Similarly, the formula ϕ can be viewed as the description of the set of object pairs $m(\phi)$, and each object pair in $m(\phi)$ as an instance of the concept given by ϕ.

4.2. Two types of rules

Knowledge derivable from an information table is commonly represented in the form of rules. Roughly speaking, rules show the connections between attributes, which are normally characterized by the problem of determining the values of one set of attributes based on the values of another set of attributes. Depending on the meanings and forms of rules, we can classify rules in many ways. A clear classification of rules is useful for an understanding of the basic tasks of machine learning and data mining.

Rules can be classified into two groups in terms of their directions, one-way and two-way connections, and further classified into two levels in terms of their applicability, local and global connections (Yao & Zhong, 1999; Yao, 2001b, 2003b). A one-way connection shows that the values of one set of attributes determine the values of another set of attributes, but does not say anything about the reverse. A two-way connection is a combination of two one-way connections, representing two different directions of connection. A local connection is characterized by a rule showing the relationship between one specific combination

of values on one set of attributes and one specific combination of values on another set of attributes. A global connection is characterized by a rule showing the relationships between all combinations of values on one set of attributes and all combinations of values on another set of attributes.

For clarity, we only consider one-way connections and the equality relation on attribute values, as was commonly done in rough sets. In this case, a local one-way connection is expressed by a rule of the form, using formulae of L_v,

$$(a, =, v_a) \Rightarrow (b, =, v_b) \tag{37}$$

where $a, b \in \mathrm{At}$, and $v_a \in V_a$, $v_b \in V_b$. It can be more conveniently expressed as, for $x \in U$,

$$I_a(x) = v_a \Rightarrow I_b(x) = v_b \tag{38}$$

The rule is commonly paraphrased as 'if the value of an object is v_a on an attribute a, then its value is v_b on another attribute b'. A global one-way connection is expressed by a rule of the form, using formulae of L_r,

$$(a, =) \Rightarrow (b, =) \tag{39}$$

where $a, b \in \mathrm{At}$, or conveniently as, for $(x, y) \in U \times U$,

$$I_a(x) = I_a(y) \Rightarrow I_b(x) = I_b(y) \tag{40}$$

That is, 'if two objects have the same value on an attribute a, then they have the same value on another attribute b'. Functional dependence in a database is an example of such global rules.

The formulation of rules using atomic formulae can be easily extended to any formulae of languages L_v and L_r. A local rule states knowledge about one object. A local one-way rule shows that, if the object has a specific value on one set of attributes, then it will have a specific value on another set of attributes. On the other hand, a global rule states knowledge about a pair of objects. A global one-way rule suggests that, if a pair of objects have the same value on one set of attributes, then they will have the same value on another set of attributes. Based on this observation, a global rule is also called a high order rule, while a local rule is called a low order rule (Yao, 2003a).

4.3. Interpretation of rough set theory in information tables

The abstract theory of rough sets can be explained by using an information table. Such an interpretation is useful for rule induction.

Let $W = \{W_1, \ldots, W_n\} \subseteq \mathrm{At}$ be a set of attributes in an information table. We form a family of elementary formulae $F_W = \{\wedge_{i=1}^n (W_i, =, w_i) | w_i \in V_{W_i}\}$ of the language L_v. For simplicity, let $V_W = V_{W_1} \times \ldots \times V_{W_n}$. We also express the family of elementary formulae by $F_W = \{W = w | w \in V_W\}$. The family of nonempty meaning sets form a partition of the universe, namely

$$\pi_W = \{m(\wedge_{i=1}^n (W_i, =, w_i)) \neq \emptyset | w_i \in V_{W_i}\} \tag{41}$$

It is referred to as the partition induced by the set of attributes W. For the same set of attributes, we can construct a formula $\wedge_{i=1}^n (W_i, =)$ of the language L_r. The meaning of the formula

$$\begin{aligned} E_W &= m(\wedge_{i=1}^n (W_i, =)) \\ &= \{(x, y) \in U \times U | I_{W_i}(x) = I_{W_i}(y), 1 \leq i \leq n\} \end{aligned} \tag{42}$$

is an equivalence relation on U. Similarly, another set of attributes $Z = \{Z_1, \ldots, Z_m\}$ defines another partition π_Z and the corresponding equivalence relation E_Z.

Rough set approximations of a single subset are relevant to the induction of local or low order rules. Consider a formula $\wedge_{j=1}^m (Z_j, =, z_j)$. In terms of attributes in W, we can obtain various local rules of the following format:

$$\wedge_{i=1}^n (W_i, =, w_i) \Rightarrow \wedge_{j=1}^m (Z_j, =, z_j) \tag{43}$$

Let $\phi_W = \wedge_{i=1}^n (W_i, =, w_i)$ and $\psi_Z = \wedge_{j=1}^m (Z_j, =, z_j)$. With respect to the three regions of rough set approximations, we can construct three classes of rules.

(I) Positive region: certain positive rules
$m(\phi_W) \subseteq m(\psi_Z)$,
$\phi_W \rightarrow \psi_Z$.

(II) Boundary region: uncertain positive rules
$m(\phi_W) \not\subseteq m(\psi_Z)$ and $m(\phi_W) \cap m(\psi_Z) \neq \varnothing$ [2],
$\phi_W \Rightarrow \psi_Z$.

(III) Negative region: certain negative rules
$m(\phi_W) \cap m(\psi_Z) = \varnothing$,
$\phi_W \rightarrow \neg\psi_Z$.

Certain rules can be considered as the degenerate cases of uncertain rules. Since certain rules can be interpreted using the logical connective $\rightarrow$, we use the same symbol. All three classes of rules express the relationship between two concepts in terms of their meaning sets. Probabilistic measures introduced earlier can be used to quantify the uncertainty of rules. For example, a measure from the rough membership function

$$\frac{|m(\phi_W) \cap m(\psi_Z)|}{|m(\phi_W)|} \tag{44}$$

can be used to measure the accuracy of one rule. Other measures associated with approximation can be used to show the characteristic of a set of rules. For instance, the measure suggested by Gediga and Düntsch (2001)

$$\begin{aligned} \underline{\gamma}_{\pi_W}(m(\psi_Z)) &= \frac{|\underline{\mathrm{apr}}_{\pi_W}(m(\psi_Z))|}{|m(\psi_Z)|} \\ &= \frac{\sum_{\phi_W \in F_W, m(\phi_W) \subseteq m(\psi_Z)} |m(\phi_W)|}{|m(\psi_Z)|} \end{aligned} \tag{45}$$

is the ratio of objects correctly classified by all certain positive rules to the objects satisfying the condition ψ_Z. Similarly, the accuracy of approximation $\alpha_{\pi_W}(m(\psi_Z))$ is the ratio of objects correctly classified by all certain positive

rules to the objects classified by both certain and uncertain positive rules.

Approximation of one partition based on another partition can be summarized by the following rule:

$$\wedge_{i=1}^{n}(W_i, =) \Rightarrow \wedge_{j=1}^{m}(Z_j, =) \qquad \text{or simply } W \Rightarrow Z \quad (46)$$

The measures $\alpha_{\pi_W}(\pi_Z)$ and $\gamma_{\pi_W}(\pi_Z)$ can be used to measure the strength of the global, high order rule.

A more detailed probabilistic and information-theoretic analysis of low and high order rules is given in the following sections (Yao & Zhong, 1999; Yao, 2003b).

4.4. Probabilistic measures for low order rules

Suppose ϕ and ψ are two formulae of the language L_v. For a rule $\phi \Rightarrow \psi$, its characteristics can be summarized by the following contingency table:

$-$	ψ	$\neg\psi$	Total
ϕ	a	b	$a+b$
$\neg\phi$	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d=\|U\|$

$$a = |m(\phi \wedge \psi)| \qquad b = |m(\phi \wedge \neg\psi)|$$
$$c = |m(\neg\phi \wedge \psi)| \qquad d = |m(\neg\phi \wedge \neg\psi)|$$

Different measures can be defined to reflect various aspects of rules.

The generality of ϕ is defined by

$$G(\phi) = \frac{|m(\phi)|}{|U|} = \frac{a+b}{|U|} \quad (47)$$

which indicates the relative size of the concept ϕ. Obviously, we have $0 \leq G(\phi) \leq 1$. A concept is more general if it covers more instances of the universe. A sub-concept has a lower generality than its super-concept. The quantity may be viewed as the probability of a randomly selected element satisfying ϕ.

The absolute support of ψ provided by ϕ is the quantity

$$\begin{aligned} \mathrm{AS}(\phi \Rightarrow \psi) &= \mathrm{AS}(\psi|\phi) \\ &= \frac{|m(\psi) \cap m(\phi)|}{|m(\phi)|} \\ &= \frac{a}{a+b} \end{aligned} \quad (48)$$

The quantity $0 \leq \mathrm{AS}(\psi|\phi) \leq 1$ states the degree to which ϕ supports ψ. It may be viewed as the conditional probability of a randomly selected element satisfying ψ given that the element satisfies ϕ. In set-theoretic terms, it is the degree to which $m(\phi)$ is included in $m(\psi)$. Clearly, $\mathrm{AS}(\psi|\phi) = 1$ if and only if $m(\phi) \neq \emptyset$ and $m(\phi) \subseteq m(\psi)$. That is, a rule with the maximum absolute support 1 is a certain rule.

The change of support of ψ provided by ϕ is defined by

$$\begin{aligned} \mathrm{CS}(\phi \Rightarrow \psi) &= \mathrm{CS}(\psi|\phi) \\ &= \mathrm{AS}(\psi|\phi) - G(\psi) \\ &= \frac{a}{a+b} - \frac{a+c}{|U|} \end{aligned} \quad (49)$$

Unlike the absolute support, the change of support varies from -1 to 1. We can consider $G(\psi)$ to be the prior probability of ψ and $\mathrm{AS}(\psi|\phi)$ the posterior probability of ψ after knowing ϕ. The difference of posterior and prior probabilities represents the change in our confidence regarding whether ϕ is actually related to ψ. For a positive value, we may say that ϕ is positively related to ψ; for a negative value, we may say that ϕ is negatively related to ψ.

The change of support relative to ψ is given by

$$\begin{aligned} \mathrm{RCS}(\psi \Rightarrow \psi) &= \frac{\mathrm{CS}(\psi|\phi)}{G(\psi)} \\ &= \frac{\mathrm{AS}(\psi|\phi)}{G(\psi)} - 1 \\ &= \frac{G(\psi \wedge \phi)}{G(\psi)G(\phi)} - 1 \\ &= \frac{|U||m(\psi) \cap m(\phi)|}{|m(\psi)||m(\phi)|} - 1 \\ &= \frac{a|U|}{(a+c)(a+b)} - 1 \end{aligned} \quad (50)$$

It is interesting to note that the first term in the relative change of support is related to the probabilistic independence of ψ and ϕ.

The generality $G(\psi)$ is related to the satisfiability of ψ by all objects in the database, and $\mathrm{AS}(\phi \Rightarrow \psi)$ is related to the satisfiability of ψ in the subset $m(\phi)$. A high $\mathrm{AS}(\phi \Rightarrow \psi)$ does not necessarily suggest a strong association between ϕ and ψ, as a concept ψ with a large $G(\psi)$ value tends to have a large $\mathrm{AS}(\phi \Rightarrow \psi)$ value. The change of support $\mathrm{CS}(\phi \Rightarrow \psi)$, or the relative change of support $\mathrm{RCS}(\phi \Rightarrow \psi)$, may be more accurate.

4.5. Information-theoretic measures for high order rules

Recall that a set of attributes W induces a partition π_W of the universe. Let

$$P(w) = P(m(W=w)) = \frac{|m(W=w)|}{|U|} \quad (51)$$

Shannon's entropy function of π_W, simply written as $H(P(W))$, is given by

$$\begin{aligned} H(P(W)) &= \boldsymbol{E}_{P(W)}[-\log P(W)] \\ &= -\sum_{w \in V_W} P(w) \log P(w) \end{aligned} \quad (52)$$

where $\boldsymbol{E}_{P(W)}[\cdot]$ denotes the expected value with respect to the probability distribution of W. For two sets of attributes

W and Z, their joint entropy is defined by

$$H(Z, W) = -\sum_{z \in V_Z} \sum_{w \in V_W} p(z, w) \log p(z, w) \tag{53}$$

The conditional entropy $H(Z|W)$ is defined as the expected value of subpopulation entropies $H(Z|w)$ with respect to the probability distribution $P(W)$:

$$\begin{aligned} H(Z|W) &= \sum_{w \in V_W} P(w) H(Z|w) \\ &= -\sum_{w \in V_W} P(w) \sum_{z \in V_Z} P(z|w) \log P(z|w) \\ &= -\sum_{z \in V_Z} \sum_{w \in V_W} P(z, w) \log P(z|w) \\ &= \boldsymbol{E}_{P(Z,W)}[-\log P(Z|W)] \end{aligned} \tag{54}$$

Conditional entropy is nonnegative and nonsymmetric, namely $H(Z|W) \geq 0$ and in general $H(Z|W) \neq H(W|Z)$. Conditional entropy can also be expressed by

$$H(Z|W) = H(Z, W) - H(W) \tag{55}$$

It measures the additional amount of information provided by Z if W is already known.

The probability $P(z)$ is the generality of the granule $m(Z = z)$. The function $-\log P(z)$ is a monotonic decreasing transformation of $P(z)$. As the expected values of $-\log P(z)$, the entropy function is related to the granularity of the partition π_Z.

The probability $P(z|w)$ is a measure for the local rule $W = w \Rightarrow Z = z$. As an expected value, the conditional entropy $H(Z|W)$ provides a measure for the global rule $W \Rightarrow Z$. It may be viewed as an inverse measure of global one-way association of two sets of attributes (Pawlak *et al.*, 1988):

$$\mathrm{IC}_1(W \Rightarrow Z) = H(Z|W) \tag{56}$$

A normalized version is given by (Pawlak *et al.*, 1988)

$$\mathrm{IC}_2(W \Rightarrow Z) = \frac{H(Z|W)}{\log |V_Z|} \tag{57}$$

For an attribute Z, conditional entropy can be used to select important attributes for discovering a one-way association $W \Rightarrow Z$. Measures IC_1 and IC_2 can be used to rank attributes in an increasing order. If one prefers to rank attributes in a decreasing order, the following corresponding direct measures of one-way association can be used:

$$C_1(W \Rightarrow Z) = \log |V_Z| - H(Z|W) \tag{58}$$

$$C_2(W \Rightarrow Z) = 1 - \frac{H(Z|W)}{\log |V_Z|} \tag{59}$$

In these measures, the attribute entropy $H(Z)$ may be used in place of $\log |V_Z|$. We obtain the following measures:

$$C_3(W \Rightarrow Z) = H(Z) - H(Z|W) = I(Z; W) \tag{60}$$

$$C_4(W \Rightarrow Z) = 1 - \frac{H(Z|W)}{H(Z)} = \frac{I(Z; W)}{H(Z)} \tag{61}$$

Measure C_3 is in fact the mutual information between W and Z. It is commonly referred to as information gain and is widely used in machine learning (Quinlan, 1986). Like the change of support for local rules, C_3 may be viewed as changes of entropy for global rules. Similarly, C_4 may be viewed as a relative change of entropy for global rules.

5. Conclusion

While nonprobabilistic studies of rough sets focus on algebraic and qualitative properties of the theory, probabilistic approaches are more practical and capture quantitative properties of the theory. The granularity of a partition can be quantified by information-theoretic measures. Existing measures of accuracy and quality of approximations can be quantified by probability-related measures. The probabilistic and information-theoretic approaches are particularly useful in rule induction, an important application of rough set theory.

Most of the measures discussed in this paper are based on simple counting of the number of elements of a set. Furthermore, we have restricted our discussion to granulations by equivalence relations or partitions. It should be pointed out that the argument can be easily extended to more general probability functions and general granulation structures.

References

DUDA, R.O. and P.E. HART (1973) *Pattern Classification and Scene Analysis*, New York: Wiley.

DÜNTSCH, I. and G.R. GEDIGA (2001) Rough information analysis, *International Journal of Intelligent Systems*, **16**, 121–147.

FAGIN, R. and J.Y. HALPERN (1991) Uncertainty, belief, and probability, *Computational Intelligence*, **7**, 160–173.

GEDIGA, G. and I. DÜNTSCH (2001) Rough approximation quality revisited, *Artificial Intelligence*, **132**, 219–234.

GRZYMALA-BUSSE, J.W. (1987) Rough-set and Dempster–Shafer approaches to knowledge acquisition under uncertainty—a comparison, Manuscript, Department of Computer Science, University of Kansas.

GRZYMALA-BUSSE, J.W. (1992) LERS—a system for learning from example based on rough sets, in *Intelligent Support: Handbook of Applications and Advances of the Rough Set Theory*, R. Slowinski (ed.), Dordrecht: Kluwer Academic, 3–18.

KLIR, G.J. and T.A. GOLGER (1988) *Fuzzy Sets, Uncertainty, and Information*, Englewood Cliffs, NJ: Prentice Hall.

LEE, T.T. (1987) An information-theoretic analysis of relational databases—part I: data dependencies and information metric, *IEEE Transactions on Software Engineering*, **SE-13**, 1049–1061.

LIN, T.Y., Y.Y. YAO and L.A. ZADEH (eds) (2002) *Data Mining, Rough Sets and Granular Computing*, Heidelberg: Physica-Verlag.

PAWLAK, Z. (1982) Rough sets, *International Journal of Computer and Information Sciences*, **11**, 341–356.

PAWLAK, Z. (1984) Rough probability, *Bulletin of Polish Academy of Sciences, Mathematics*, **32**, 607–615.

PAWLAK, Z. (1991) *Rough Sets, Theoretical Aspects of Reasoning about Data*, Dordrecht: Kluwer Academic.
PAWLAK, Z. and A. SKOWRON (1994) Rough membership functions, in *Advances in the Dempster–Shafer Theory of Evidence*, R.R. Yager, M. Fedrizzi and J. Kacprzyk (eds), New York: Wiley, 251–271.
PAWLAK, Z., S.K.M. WONG and W. ZIARKO (1988) Rough sets: probabilistic versus deterministic approach, *International Journal of Man–Machine Studies*, **29**, 81–95.
QUINLAN, J.R. (1986) Induction of decision trees, *Machine Learning*, **1**, 81–106.
SKOWRON, A. (1989) The relationship between the rough set theory and evidence theory, *Bulletin of Polish Academy of Sciences, Mathematics*, **37**, 87–90.
SKOWRON, A. (1990) The rough set theory and evidence theory, *Fundamenta Informaticae*, **13**, 245–262.
SKOWRON, A. and J. GRZYMALA-BUSSE (1994) From rough set theory to evidence theory, in *Advances in the Dempster–Shafer Theory of Evidence*, R.R. Yager, M. Fedrizzi and J. Kacprzyk (eds), New York: Wiley, 193–236.
TSUMOTO, S. (1998) Automated extraction of medical expert system rules from clinical databases on rough set theory, *Information Sciences*, **112**, 67–84.
WONG, S.K.M. and P.J. LINGRAS (1989) The compatibility view of Shafer–Dempster theory using the concept of rough set, *Methodologies of Intelligent Systems*, **4**, 33–42.
WONG, S.K.M. and W. ZIARKO (1986) Algorithm for inductive learning, *Bulletin of the Polish Academy of Sciences, Technical Sciences*, **34**, 271–276.
WONG, S.K.M. and W. ZIARKO (1987) Comparison of the probabilistic approximate classification and the fuzzy set model, *Fuzzy Sets and Systems*, **21**, 357–362.
YAO, J.T. and Y.Y. YAO (2002) Induction of classification rules by granular computing, *Proceedings of the Third International Conference on Rough Sets and Current Trends in Computing*, LNAI 2475, Berlin: Springer, 331–338.
YAO, Y.Y. (1996) Two views of the theory of rough sets in finite universes, *International Journal of Approximation Reasoning*, **15**, 291–317.
YAO, Y.Y. (2000) Granular computing: basic issues and possible solutions, *Proceedings of the 5th Joint Conference on Information Sciences*, Association for Intelligent Machinery, 186–189.
YAO, Y.Y. (2001a) Information granulation and rough set approximation, *International Journal of Intelligent Systems*, **16**, 87–104.
YAO, Y.Y. (2001b) Modeling data mining with granular computing, *Proceedings of the 25th Annual International Computer Software and Applications Conference*, 638–643.
YAO, Y.Y. (2003a) Mining high order decision rules, in *Rough Set Theory and Granular Computing*, M. Inuiguchi, S. Hirano and S. Tsumoto (eds), Berlin: Springer, 125–135.
YAO, Y.Y. (2003b) Information-theoretic measures for knowledge discovery and data mining, in *Entropy Measures, Maximum Entropy and Emerging Applications*, Karmeshu (ed.), Berlin: Springer, 115–136.
YAO, Y.Y. and P.J. LINGRAS (1998) Interpretations of belief functions in the theory of rough sets, *Information Sciences*, **104**, 81–106.
YAO, Y.Y. and S.K.M. WONG (1992) A decision theoretic framework for approximating concepts, *International Journal of Man–Machine Studies*, **37**, 793–809.
YAO, Y.Y. and N. ZHONG (1999) An analysis of quantitative measures associated with rules, *Proceedings of the Third Pacific-Asia Conference on Knowledge Discovery and Data Mining*, LNAI 1574, Berlin: Springer, 479–488.
YAO, Y.Y., S.K.M. WONG and P.J. LINGRAS (1990) A decision-theoretic rough set model, in *Methodologies for Intelligent Systems*, Vol. 5, Z.W. Ras, M. Zemankova and M.L. Emrich (eds), New York: North-Holland, 17–24.
ZADEH, L.A. (1979) Fuzzy sets and information granularity, in *Advances in Fuzzy Set Theory and Applications*, N. Gupta, R. Ragade and R. Yager (eds), Amsterdam: North-Holland, 3–18.
ZADEH, L.A. (1997) Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, *Fuzzy Sets and Systems*, **19**, 111–127.
ZHANG, B. and L. ZHANG (1992) *Theory and Applications of Problem Solving*, Amsterdam: North-Holland.
ZIARKO, W. (1993) Variable precision rough set model, *Journal of Computer and System Sciences*, **46**, 39–59.

The author

Yiyu Yao

Yiyu Yao is a professor of computer science in the Department of Computer Science, University of Regina, Regina, Saskatchewan, Canada. His research interests are information retrieval, Web intelligence, data mining, fuzzy sets, rough sets and granular computing. He received a PhD degree from the University of Regina, Canada. He has published over 100 journal and conference papers. He is a member of the editorial boards of the *Web Intelligence and Agent Systems* journal (IOS Press). He has served and is serving as a program co-chair of three international conferences, and as a program committee member in over 20 international conferences. He is a member of IEEE and ACM.