Generalization of Rough Sets Using Relationships Between Attribute Values

Y.Y. Yao

Department of Computer Science, Lakehead University
Thunder Bay, Ontario, Canada P7B 5E1
E-mail: yyao@flash.lakeheadu.ca

S.K.M. Wong

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

abstract The notion of rough sets is generalized by using an arbitrary binary relation on attribute values in information systems, instead of the trivial equality relation. The relation, not necessarily equivalence, between objects are defined in terms of relationships between attribute values. The adoption of other types of relations enables us to define various classes of rough set models. This study provides a basis for non-standard rough sets. The results are complementary to that of investigations based on modal logic.

1 Introduction

In the Pawlak rough set model, an equivalence relation on the universe of objects is defined based on their attribute values. In particular, this equivalence relation is constructed based on the trivial equality relation = on attribute values. Although the use of equality relation has produced many interesting and useful results, it may not be sufficient in some application (Lin, 1988; Zakowski, 1983).

To resolve the limitations imposed by an equivalence relation, many proposals have been made. Zakowski (1983) suggested that one may use a compatibility relation, i.e., a reflexive and symmetric relation, instead of an equivalence relation on the universe. Wybraniec-Skardowska (1989) introduced different rough-set models based on various types of binary relations. Based on the results from modal logic, Yao and Lin (1995) examined non-standard rough set models induced by various binary relations. Unlike the Pawlak rough set model, a crucial gap in these studies is that the process of deriving the relation on the universe was not discussed in the context of information systems. Many authors have attempted to address this problem. The notion of weak discernibility relation adopted by Vakarelov (1991) is a compatibility relation, which is interpreted using the intersection of subsets of attribute values. Instead taking a single value, an object takes a subset as its value. Two objects are compatible if they share some common attribute values, i.e., nonempty intersection. On the other hand, Lin (1988) discussed relation on objects based on single values in the neighborhood systems. In terms of their neighborhood systems, two objects are related if their attributed values are related. Like the Pawlak rough set model, in both methods, the relation on objects are defined by the relationship between their attribute values, which in general is not an equivalence relation.

Based on the above studies, this paper proposes a systematic method for the construction of a binary relation on objects using an arbitrary binary relation on their attribute values. In the context of information systems, different non-standard rough set models are derived from different types of binary relations on attribute values. This study complete investigations based on modal logic by providing a plausible explanation of relations on the universe of objects.

2 Pawlak Rough Sets

Following Lipski (1981), Orlowska (1985), Pawlak (1981), Vakarelov (1991), and Yao and Noroozi (1994),

we define a set-based information system to be a quadruple,

$$S = (U, At, \{V_a \mid a \in A\}, \{f_a \mid a \in A\}),$$

where

U is a nonempty set of objects,
At is a nonempty set of attributes,
V_a is a nonempty set of values of $a \in A$,
$f_a : U \longrightarrow 2^{V_a}$ is an information function.

The notion of information systems provides a convenient tool for the representation of objects in terms of their attribute values. If all information functions map an object to only singleton subsets of attribute values, we obtain a degenerate set-based information system commonly used in the Pawlak rough-set model. For clarity, we only consider this kind of information systems. In this case, information functions can be expressed as $f_a : U \longrightarrow V_a$.

With the above information, we can describe relationships between objects through their attribute values. With respect to an attribute $a \in At$, an relation $\Re_a$ is given by: for $o, o' \in U$,

$$o\Re_a o' \Longleftrightarrow f_a(o) = f_a(o'). \tag{1}$$

Obviously, $\Re_a$ is an equivalence relation. The reflexivity, symmetry and transitivity of $\Re_a$ follow trivially from the properties of the relation $=$ between attribute values. With the defined equivalence relation, two objects are considered to be indiscernible, in the view of attribute a, if and only if they have exactly the *same* value on a. This definition can be extended to any subset $A \subseteq At$ as follows:

$$o\Re_A o' \Longleftrightarrow (\forall a \in A) f_a(o) = f_a(o'). \tag{2}$$

Relation $\Re_A$ is an equivalence relation (Orlowska, 1985). That is, in terms of all attributes in A, o and o' are indiscernible.

The pair $apr_A = (U, \Re_A)$ is referred to as an approximation space. For any element $o \in U$, we can construct its equivalence class, i.e., all elements $\Re_A$ related to o:

$$r_A(o) = \{o' \mid o\Re_A o'\}. \tag{3}$$

Using these equivalence classes, for any subset $X \subseteq U$ we have the following lower and upper approximations:

$$\begin{aligned} \underline{apr}_A(X) &= \{o \mid r_A(o) \subseteq X\}, \\ \overline{apr}_A(X) &= \{o \mid r_A(o) \cap X \neq \emptyset\}. \end{aligned} \tag{4}$$

The pair $(\underline{apr}_A(X), \overline{apr}_A(X))$ is referred to as the rough set of X in the approximation space apr_A.

For any subsets $X, Y \subseteq U$, the lower approximation satisfies properties:

(L1) $\underline{apr}_A(X) = \ \sim\overline{apr}_A(\sim X)$,
(L2) $\underline{apr}_A(U) = U$,
(L3) $\underline{apr}_A(X \cap Y) = \underline{apr}_A(X) \cap \underline{apr}_A(Y)$,
(L4) $\underline{apr}_A(X \cup Y) \supseteq \underline{apr}_A(X) \cup \underline{apr}_A(Y)$,
(L5) $X \subseteq Y \Longrightarrow \underline{apr}_A(X) \subseteq \underline{apr}_A(Y)$,
(L6) $\underline{apr}_A(X) \subseteq \overline{apr}_A(X)$,
(L7) $\underline{apr}_A(X) \subseteq X$,
(L8) $X \subseteq \underline{apr}_A(\overline{apr}_A(X))$,
(L9) $\underline{apr}_A(X) \subseteq \underline{apr}_A(\underline{apr}_A(X))$,
(L10) $\overline{apr}_A(X) \subseteq \underline{apr}_A(\overline{apr}_A(X))$,

where $\sim A = U - A$ denotes the set complement of A. Similar properties can be established for the upper approximation.

3 Generalized Rough Sets

The Pawlak rough set model can be easily generalized by considering any type of binary relations on attribute values, instead of the trivial equality relation $=$. Suppose R_a is a binary relation on the values of an attribute $a \in At$. By extending equation (1), for $a \in At$ we define a binary relation on U:

$$o\Re_a o' \Longleftrightarrow f_a(o) R_a f_a(o'). \tag{5}$$

Similarly, by extending equation (2), for $A \subseteq At$ we define a relation on U:

$$\begin{aligned} o\Re_A o' &\iff (\forall a \in A) f_a(o) R_a f_a(o') \\ &\iff (\forall a \in A) o\Re_a o'. \end{aligned} \tag{6}$$

An object o is related to another object o', based on only attribute a, if their values on a are related. With respect to a subset A of attributes, o is related to o' their values are related for every attributes in A. When all relations R_a are chosen to be $=$, the proposed definition reduced to the definition in the Pawlak rough set model.

Like the Pawlak rough set model, the empty set $\emptyset$ produces the coarsest relation, i.e., $\Re_\emptyset = U \times U$, where $\times$ denotes the Cartesian product of sets. If the entire attribute set is used, one obtains the finest relation $\Re_{At}$. Moreover, if each object is described by an unique description, $\Re_{At}$ becomes the identity relation. The algebra $(\{\Re_A\}_{A \subseteq At}, \cap)$ is a lower semilattice with the zero element $\Re_{At}$ (Orlowska, 1985).

Theorem 1 Suppose R_a is a binary relation on V_a, and $\Re_A$ a binary relation on U defined by equation (6). Independent of the properties of R_a, the following properties hold: for $A, B \subseteq At$,

$$\text{(I1)} \qquad \Re_{(A \cup B)} = \Re_A \cap \Re_B,$$
$$\text{(I2)} \qquad A \subseteq B = \Re_B \subseteq \Re_A.$$

Property (I2) trivially follows from (I1), which implies that a finer relation is obtained by adding more attributes. However, $\Re_{(A \cap B)}$ cannot be obtained from $\Re_A$ and $\Re_B$.

Consider the following types of binary relations: serial, reflexive, symmetric, transitive and Euclidean relations.[1] By its definition, $\Re_a$ preserves properties of R_a.

Lemma 1 Suppose R_a is a binary relation on V_a, and $\Re_a$ a binary relation on U defined by equation (5). Then,

a). R_a is serial $\Longrightarrow \Re_a$ is serial;
b). R_a is reflexive $\Longrightarrow \Re_a$ is reflexive;
c). R_a is symmetric $\Longrightarrow \Re_a$ is symmetric;
d). R_a is transitive $\Longrightarrow \Re_a$ is transitive;
e). R_a is Euclidean $\Longrightarrow \Re_a$ is Euclidean.

Properties of R_a are sufficient but not necessary conditions for the corresponding properties of $\Re_a$. Assume further that information systems obey the condition:

$$(\forall v \in V_a)(\exists o \in U) f_a(o) = v. \tag{7}$$

That is, for every attribute value $v \in V_a$, there exists at least one object whose value for attribute a is v. Under this assumption, properties of R_a are both necessary and sufficient conditions. The single implication in Lemma 1 becomes double implication. For instance, $\Re_a$ is reflexive if and only if R_a is reflexive.

By property (I1), relation $\Re_A$ can be expressed in terms of $\Re_a$'s:

$$\Re_A = \bigcap_{a \in A} \Re_a. \tag{8}$$

Combining this with the results of Lemma 1, one may conclude that R_a's determine the properties of $\Re_A$.

[1] A relation R on a set X is a serial relation if for all $x \in X$ there exists a $y \in X$ such that $x \Re y$. A relation is a reflexive relation if for all $x \in X$ the relationship $x \Re x$ holds. A relation is symmetric relation if for all $x, y \in X$, $x \Re y$ implies $y \Re x$ holds. A relation is a transitive relation if for three elements $x, y, z \in X$, $x \Re y$ and $y \Re z$ imply $x \Re z$. A relation is Euclidean if for all $x, y, z \in X$, $x \Re y$ and $x \Re z$ imply $y \Re z$.

Theorem 2 Suppose R_a is a binary relation on V_a, and $\Re_A$ a binary relation on U defined by equation (6). Then,

b′). $(\forall a \in A) R_a$ is reflexive $\Longrightarrow \Re_A$ is reflexive;
c′). $(\forall a \in A) R_a$ is symmetric $\Longrightarrow \Re_A$ is symmetric;
d′). $(\forall a \in A) R_a$ is transitive $\Longrightarrow \Re_A$ is transitive;
e′). $(\forall a \in A) R_a$ is Euclidean $\Longrightarrow \Re_A$ is Euclidean.

In this theorem, all R_a's have the same property in order to derive the corresponding property of the relation $\Re_A$. As noted for Lemma 1, they are only sufficient but not necessary conditions. If A contains more than one attribute, the assumption (7) no longer implies that all those conditions are also necessary. For example, under assumption (7), $\Re_A$ is reflexive if and only if for all $a \in A$, R_a is reflexive. In contrast, the symmetry of $\Re_A$ does not require that for all $a \in A$, R_a is symmetric. There is no counterpart of (a) in the Lemma 1. In general, the condition, $(\forall a \in A) R_a$ is serial, is neither a sufficient nor a necessary condition for $\Re_A$ to be serial. Under assumption (7), it becomes a necessary condition but still not sufficient.

The pair $apr_A = (U, \Re)$ is called a generalized approximation space, where the relation $\Re$ is not necessarily an equivalence relation. Using $\Re_A$, one can immediately apply equations (3) and (4) to define $\Re_A$-related elements and generalized rough sets. That is, the following definitions can be used:

$$\begin{aligned} r_A(o) &= \{o' \mid o \Re_A o'\}, \\ \underline{apr}_A(X) &= \{o \mid r_A(o) \subseteq X\}, \\ \overline{apr}_A(X) &= \{o \mid r_A(o) \cap X \neq \emptyset\}. \end{aligned} \tag{9}$$

The pair $(\underline{apr}_A(X), \overline{apr}_A(X))$ is referred to as the generalized rough set of X in the generalized approximation space apr_A. Properties of generalized lower and upper approximations are determined by properties of $\Re_A$, which in turn are determined by the properties of R_a's.

Theorem 3 Suppose R_a is a binary relation on V_a, and $\Re_A$ a binary relation on U defined by equation (6). Independent of the properties of R_a, (L1)-(L5) hold for the generalized lower approximation.

Properties (L1)-(L5) directly follow from the definition of generalized rough sets. The condition for property (L6) is that $\Re_A$ is a serial relation. Conditions for properties (L7)-(L10) are that $\Re_A$ are reflexive, symmetric, transitive and Euclidean relations, respectively (Yao and Lin, 1995). By Theorem 2, such conditions can be expressed by the properties of R_a.

Theorem 4 Suppose R_a is a binary relation on V_a, and $\Re_A$ a binary relation on U defined by equation (6). Then,

a). $(\forall a \in A)R_a$ is reflexive $\Longrightarrow$ (L7) holds;

b). $(\forall a \in A)R_a$ is symmetric $\Longrightarrow$ (L8) holds;

c). $(\forall a \in A)R_a$ is transitive $\Longrightarrow$ (L9) holds;

d). $(\forall a \in A)R_a$ is Euclidean $\Longrightarrow$ (L10) holds.

Yao and Lin (1995) examined the classification of rough set models using properties such as (L6)-(L10), based on the results from modal logic. The above discussion clearly provides a concrete and systematic method for constructing relations on the universe. An implication of Theorem 4 is that we can use properties of the relationship between attribute values to describe various rough set models.

The set of $\Re_A$-related objects, $r_A(o) = \{o' \mid o\Re_A o'\}$, can be regarded as a neighborhood of o. Likewise, the set of R_a-related values, $r_a(v) = \{v' \mid vR_a v'\}$, can be viewed as a neighborhood of v (Lin, 1988). By definition, a neighborhood of objects is defined according to neighborhoods of its attribute values:

$$\begin{aligned} r_A(o) &= \{o' \mid o\Re_A o'\} \\ &= \bigcap_{a\in A} \{o' \mid o\Re_a o'\} \\ &= \bigcap_{a\in A} \{o' \mid f_a(o)R_a f_a(o')\} \\ &= \bigcap_{a\in A} \{o' \mid f_a(o') \in r_a(f_a(o))\}. \end{aligned} \quad (10)$$

It suggests that the results of this study can also be applied to approximate retrieval in information systems (Lin, 1988).

4 Conclusion

In this paper, we have presented a systematic method for the construction of relations on the universe of objects based on their relationships on their properties. It have clearly demonstrated that the properties of relations on attribute values determine the properties of induced relations on the universe of objects. Since the relationships between attribute values are not restricted the trivial equality relation, the proposed model generalizes Pawlak rough sets constructed from an equivalence relation. This enables us to define various classes of rough set models in the context of information systems. This study provides a basis for non-standard rough sets. The results are complementary to that of investigations based on modal logic, and can be applied to approximate retrieval in information systems.

References

Lin, T.Y. (1988). Neighborhood systems and approximation in relational databases and knowledge bases. *Proceedings of the 4th International Symposium on Methodologies of Intelligent Systems.*

Lipski, W. Jr. (1981). On databases with incomplete information. *Journal of the ACM*, **28**, 41-70.

Orlowska, E. (1985). Logic of indiscernibility relations. *Lectures Notes in Computer Science*, vol. 208, Springer-Verlag, Berlin, 177-186.

Pawlak, Z. (1981). Information systems – theoretical foundations. *Information Systems*, **6**, 205-218.

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

Vakarelov, D. (1991). A modal logic for similarity relations in Pawlak knowledge representation systems. *Fundamenta Informaticae*, **XV**, 61-79.

Wybraniec-Skardowska, U. (1989). On a generalization of approximation space. *Bulletin of the Polish Academy of Sciences*, **37**, 51-61.

Yao, Y.Y. and Lin, T.Y. (1995). Generalization of rough sets using modal logic. Manuscript.

Yao, Y.Y. and Noroozi, S. (1994). A unified model for set-based computations. *Proceedings of the 3rd International Workshop on Rough Sets and Soft Computing*, 236-243.

Zakowski, W. (1983). Approximations in the space (U, Π). *Demonstratio Mathematica*, **XVI**, 761-769.