Information-Based Algorithm for Reduction of Knowledge

Miao Duoqian Wang Jue
Institute of Automation
Chinese Academy of Sciences
Beijing 100080
P.R.China

Abstract- In Rough Set(RS) theory, it has been proved that finding the minimal reduct of an information system is an NP-complete problem. Because of this, it is hard to obtain the set of the most concise rules by existing algorithm in RS for reduction of knowledge. In this paper, an information-based algorithm for reduction of knowledge is proposed, and its time complexity is analyzed. Through an example, we show that the proposed algorithm is effective for dealing with relatively large-scale databases.

I. INTRODUCTION

Rough Set(RS) theory introduced by Prof. Z.Pawlak in 1982[3] is a new mathematical tool to reason about vagueness and uncertainty. It also provides techniques to reduce knowledge in databases, by which the irrelevant or superfluous knowledge(attributes) can be eliminated according to the learning task without losing essential information about the original data in the databases. As a result of the reduction of knowledge, a set of concise and meaningful rules are produced.

As is well known that an information system may usually have more than one reduct. This means the set of rules derived from reduction of knowledge is not unique. In practice, it is always hoped to obtain the set of the most concise rules. Therefore, people have been attempting to find the minimal reduct of information systems, which means that the number of attributes contained in the reduct is minimal. Unfortunately, it has been proven that finding the minimal reduct of an information system is an NP-complete problem[4]. Because of this, it is hard to obtain the set of the most concise rules by existing algorithm in RS for reduction of knowledge.

In this paper, we mainly discuss the reduction of knowledge in information system $S=<U,A,V,f>$. We consider equivalence relations (knowledge or attributes) in S as random variables defined over U. Entropy of knowledge is defined, so it can be used as the significance of attributes. Then we define reduction of knowledge from the view of information. Based on these definitions, we propose a bottom-up strategy for constructing the reduct by using core as the starting point. The complexity of our algorithm is $O(N^2)$, and in most cases, this algorithm can find the minimal reduct. An illustrative example for our algorithm is also given.

II. NEW DEFINITION OF REDUCTION OF KNOWLEDGE------FROM THE INFORMATIONAL PROSPECTIVE

We assume that knowledge discussed in this paper is represented by information system. An information system $S=<U,A,V,f>$ can be conveniently represented by a table, for example, table I in section 4.

According to [2], any subsets $P,Q\subseteq A$ can be considered as random variables defined over U, and their probability distributions are defined as follows respectively:

Let
$$X=U/P=\{X_1,X_2,\cdots,X_n\}$$
$$Y=U/Q=\{Y_1,Y_2,\cdots,Y_m\},$$

then probability distributions of P and Q are

$$[X;p]=\begin{bmatrix} X_1 & X_2 & \cdots & X_n \\ p(X_1) & p(X_2) & \cdots & p(X_n) \end{bmatrix}$$

and
$$[Y;p]=\begin{bmatrix} Y_1 & Y_2 & \cdots & Y_m \\ p(Y_1) & p(Y_2) & \cdots & p(Y_m) \end{bmatrix};$$

and their joint probability distribution is

0-7803-4253-4/97/$10.00

$$|XY; p| = \begin{bmatrix} X_1 \cap Y_1 \cdots X_i \cap Y_j \cdots X_n \cap Y_m \\ p(X_1 Y_1) \cdots p(X_i Y_j) \cdots p(X_n Y_m) \end{bmatrix},$$

where, $p(X_i) = \dfrac{card X_i}{card U}$, i=1,2,⋯,n;

$p(Y_j) = \dfrac{card Y_j}{card U}$, j=1,2,⋯,m;

$p(X_i Y_j) = \dfrac{card(X_i \cap Y_j)}{card U}$, i=1,2,⋯ n ; j=1,2,⋯ m;

$U/_P$ means the family of all equivalence classes of P over U, $card$ denotes the cardinality of the set.

According to information theory[1], the entropy of P and the conditional entropy of Q under P are defined as follows:

$$H(P) = -\sum_{i=1}^{n} p(X_i) \log p(X_i),$$

and $H(Q|P) = -\sum_{i=1}^{n} p(X_i) \sum_{j=1}^{m} p(Y_j|X_i) \log p(Y_j|X_i)$.

In RS theory[4], a subset $R \subseteq A$ is called a reduct of S, if $IND(R) = IND(A)$ and R is independent, where $IND(A)$ is the indiscernibility relation over A. A reduct of knowledge is its essential part, which suffices for defining all basic concepts occurring in the considered knowledge.

Theorem 1. Given an information system $S =< U, A, V, f >$, and let $R \subseteq A$.

If 1) $H(R) = H(A)$;

2) $H(a|R) > 0$, for every $a \in R$.

then R is called a reduct of S.

Proof. Follows from Lemmas 1 and 2 below.

Lemma 1. Given an information system $S =< U, A, V, f >$, and let $R, Q \subseteq A$.

Then $IND(R) = IND(Q)$ if and only if $H(R) = H(Q)$ and $R \subseteq Q$ (or $Q \subseteq R$).

Proof. See theorem 3 in [2].

Lemma 2. Given an information system $S =< U, A, V, f >$, and let $a \in A$. Then a is indispensible in A if and only if $H(a|A) > 0$, for every $a \in A$.

Proof. See theorem 4 in [2].

Theorem 1 provides the definition of reduct of knowledge from the informational prospective, which is the theoretical foundation for our algorithm given in the next section.

III. INFORMATION-BASED ALGORITHM FOR REDUCTION OF KNOWLEDGE

In this section, a bottom-up strategy for constructing reduct is presented. Because core is the common part of all reducts, the core can be used as the starting point of the algorithm. According to Theorem 1 above, the entropy of attributes can be used to select the attributes to be added to the core. We are mainly interested in the minimal reduct. Finding it, however, is in general an NP-complete problem, so we are satisfied with the following approximate procedure.

Input: An information system $S =< U, A, V, f >$.

Output: One reduct R of S.

Step 1: Compute the entropy $H(A)$ of the original set of attributes A.

Step 2: Compute the core C. For the core, the entropy $H(C) \leq H(A)$ by definition.

In particular, the core may be empty and then $H(C) = 0$.

Step 3: Set $R := C$. For the set of attributes A repeat:

(a) Compute conditional entropy $H(a|R)$ for every attribute $a \in A - R$.

(b) Choose attribute a which satisfys equation $H(a|R) = \max_{a_i \in A-R} \{H(a_i|R)\}$, and $R := R \cup \{a\}$.

(c) If $H(R) = H(A)$, then stop; otherwise go to step (a).

Step 4: The result R constitutes a reduct.

Remark: In most cases, the reduct constructed by our procedure is possibly, but not necessarily, the minimal reduct.

Time complexity: Generally speaking, the number of attributes in an information system is small, so it isn't taken into account in the complexity. In the worst case, we can find a relatively minimal reduct in $O(N^2)$, where N is the number of objects in U.

IV. AN ILLUSTRATIVE EXAMPLE

In this section, an illustrative example for our algorithm is presented. Let us consider the information system S about car[4], see the table 1. $S=<U,A,V,f>$, where U={1,2,...,21}, A={make-model, fuel, disp, weight, cyl, power, t, comp, tran, mileage}.

1. The entropy of the original set of attributes A is H(A)=4.1066.

2. The computed core C=CORE(A)={power, tran, t, comp}; and H(C)=3.3087.

3. Set R:=C. $H(R) < H(A)$, so R does not constitute a reduct. Therefore, compute the conditional entropy H(a|R) for every attribute $a \in A-R$:

Attribute a	weight	mileage	cyl	fuel	make-model	disp
H(a|R)	0.5714	0.5714	0.3860	0.3450	0.3450	0.0952

It can be seen that there are two attributes(i.e. weight and mileage) with the same maximum H(a|R). In this case, we should choose the one which has the least number of combinations of values with R. In this example, the attribute "weight" is selected. Hence, the new R={power, tran, t, comp, weight}, and the new H(R)=3.8801.

Next, using the same method as before, we find the attribute "fuel" has the maximum H(a|R). This time the new R={ power, tran, t, comp, weight, fuel}, and the new H(R)=4.1066. Because H(R)=H(A), the procedure stops. The last set of attributes R is one reduct of the information system CAR(see table 2 below).

The power of classification of Table 2 is the same as that of Table 1, but Table 2 is more simple than Table 1.

TABLE I. THE INFORMATION SYSTEM ABOUT CAR

make-model	fuel	disp	weight	cyl	power	t	comp	tran	mileage
USA	EFI	med	medium	6	high	y	high	manu	medium
USA	EFI	med	medium	6	high	n	medium	manu	medium
USA	EFI	med	medium	6	high	n	medium	auto	medium
USA	EFI	med	medium	6	medi	n	medium	manu	medium
USA	EFI	med	medium	6	high	n	high	manu	medium
USA	EFI	med	light	4	high	y	high	manu	high
USA	EFI	sma	medium	6	medi	n	high	manu	high
USA	EFI	med	medium	6	high	n	high	manu	medium
USA	EFI	med	heavy	6	high	y	high	auto	low
USA	EFI	med	medium	6	medi	n	medium	manu	medium
USA	EFI	med	heavy	4	high	n	medium	manu	low
Japan	2-BBL	sma	light	4	low	n	high	manu	high
USA	2-BBL	sma	medium	4	medi	n	high	auto	medium
Japan	EFI	sma	medium	4	low	n	high	manu	high
Japan	EFI	med	light	4	medi	n	medium	manu	high
USA	2-BBL	med	medium	4	medi	n	medium	manu	medium
Japan	EFI	sma	medium	4	high	y	high	auto	high
Japan	2-BBL	sma	medium	4	low	n	medium	manu	high
USA	EFI	med	medium	4	high	y	medium	manu	medium
USA	EFI	med	medium	6	high	n	medium	manu	medium
Japan	EFI	sma	medium	4	medi	n	high	auto	high

TABLE II. THE REDUCED INFORMATION SYSTEM ABOUT CAR

fuel	weight	power	t	comp	tran
EFI	medium	high	y	high	manu
EFI	medium	high	n	medium	manu
EFI	medium	high	n	medium	auto
EFI	medium	medi	n	medium	manu
EFI	medium	high	n	high	manu
EFI	light	high	y	high	manu
EFI	medium	medi	n	high	manu
EFI	heavy	high	y	high	auto
EFI	heavy	high	n	medium	manu
2-BBL	light	low	n	high	manu
2-BBL	medium	medi	n	high	auto
EFI	medium	low	n	high	manu
EFI	light	medi	n	medium	manu
2-BBL	medium	medi	n	medium	manu
EFI	medium	high	y	high	auto
2-BBL	medium	low	n	medium	manu
EFI	medium	high	y	medium	manu
EFI	medium	medi	n	high	auto

V. CONCLUSION

In this paper, we define reduction of knowledge from the view of information, and use information entropy of attributes to define the significance of the attributes. A bottom-up strategy which is based on information is presented for constructing an approximately minimal reduct. Through experiments, it is shown that in most cases, this algorithm can find the minimal reduct of given information systems.

VI. REFERENCES

[1] Jian,D. and Qian,Y.M., Information Theory and Coding(in Chinese), The Science and Technology Press of China, 1992.

[2] Miao, D.Q. and Wang,J., " An Information Entropy Based Algorithm for Reduction of Knowledge(in Chinese)", Technical Report, Inst. of Automation, Chinese Academy of Sciences, 1996.

[3] Pawlak,Z., Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer ,1991.

[4] Ziarko,W., "The Discovery, Analysis and Representation of Data Dependencies in Databases", Knowledge Discovery in Databases, AAAI MIT Press, Cambridge, MA, 1993, 213-228.