***I.J. Education and Management Engineering* 2011, 1, 6-11**
Published Online July 2011 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2011.01.02

Available online at http://www.mecs-press.net/ijeme

Research on Feature Selection Algorithm in Rough Set Based on Information Entropy

Guijuan Song

Department of Computer Science and Technology Dalian Neusoft Institute of Information Dalian, China

Abstract

Rough set theory is an effective approach to imprecision, vagueness, and incompleteness in classification analysis and knowledge discovery .Attribute reduction is a key problem for rough set theory. While computing reduction according to the definitions is a typical NP problem. In this paper, basic concept of rough set theory is presented, one heuristic algorithm for attribution reduction based on conditional entropy is proposed. The actual application shows that the method is feasible and effective.

Index Terms: rough set; attribute reduction;decision table;discernibility matrix;information entropy

1. Introduction

The theory of rough set proposed by Pawlak [1] in 1982 is a new method for data processing based on equivalence relation, and it has been successfully applied in such artificial intelligence fields as machine learning, pattern recognition, decision analysis, process control, knowledge discovery in databases and expert systems. The main idea of rough set theory is attribute reduction. It is well known that an information system or a decision table may usually have more than one reduction. It has been proven that finding the minimal reduction of information systems or decision tables is an NP-complete problem. Usually heuristic algorithm is used for reduction of attributes. In this paper, using discernibility matrix introduced by Skowron and Rauszer and conditional entropy, we propose a heuristic algorithm. By constructing an example, we show that the proposed algorithm is feasible.

The rest of the article is organized as follows. Section 2 presents the fundamentals of Pawlak's rough sets. In Section 3, we propose a heuristic algorithm. In Section 4, we show some example on fourteen public data sets. Section 5 presents conclusions.

Corresponding author:
E-mail address: songguijuan@neusoft.edu.cn

2. Preliminaries

2.1. Information System and Decision Table

Rough sets have been employed to remove redundant conditional attributes from discrete-valued data sets while retaining their information content. Here are some basic concepts.

Let $I=(U;\ A)$ be an information system, where U is a non-empty set of finite objects (the universe of discourse); A is a non-empty finite set of attributes such that $a: U\rightarrow Va$, $\forall a\in A$; Va being the value set of attribute a. In a decision system, $A=\{C\cup D\}$ where C is the set of conditional attributes and D is the set of decision attributes. With any $P\subseteq A$ there is an associated equivalence relation IND(P) [1]

$$IND(P)=\{(X,Y)\in U^2 \mid \forall a\in P, a(x)=a(y)\}. \quad (1)$$

The partition of U, generated by IND(P) is denoted U/P and can be calculated as follows:

$$U/P=\otimes\{a\in P: U/IND(\{a\})\}, \quad (2)$$

where

$$A\otimes B=\{X\cap Y: \forall X\in A, \forall Y\in B, X\cap Y\neq\emptyset\}.$$

If $(x,y)\in IND(P)$, then x and y are indiscernible by attributes from P. The equivalence classes of the P-indiscernibility relation are denoted $[x]_P$. Let $X\subseteq U$, the *P-lower* approximation[2] of a set can now be defined as

$$\underline{PX}=\{x\mid [x]_P\subseteq X\} \quad (3)$$

Let P and Q be equivalence relations over U, then the positive region [3] can be defined as

$$POS_P(Q)=\bigcup_{X\in U/Q}\underline{P}X\ . \quad (4)$$

In terms of classification, the positive region contains all objects of U that can be classified to classes of U/Q using the knowledge in attributes P.

Let U be a universe, P, Q denote a family of equivalence relations on the universe. Then P, Q may be considered as random variables on the σ-algebra that is composed of the subsets of the universe U. Let X, Y be two partitions of the universe induced respectively by P, Q, where

$X=U/\mathrm{IND}(P)=\{X_1, X_2,\ldots, X_n\}$,

$Y=U/\mathrm{IND}(Q)=\{Y_1, Y_2,\ldots, Y_m\}$.

Then probability distributions of X,Y are defined respectively by:

$$[X:\ \mathrm{p}]=\begin{bmatrix} X_1 & X_2 & \ldots & X_n \\ p(X_1) & p(X_2) & \ldots & p(X_n) \end{bmatrix}, \quad (5)$$

$$[Y:\ \mathrm{p}]=\begin{bmatrix} Y_1 & Y_2 & \ldots & Y_m \\ p(Y_1) & p(Y_2) & \ldots & p(Y_m) \end{bmatrix}. \quad (6)$$

where

$$p(X_i) = \frac{card(X_i)}{card(U)}, i = 1, 2, ..., n;$$

$$p(X_i) = \frac{card(Y_j)}{card(U)}, j = 1, 2, ..., m.$$

The "card(.)" denotes the cardinality of a set.

Having defined the probability distribution of knowledge[4], we can give the definitions of information entropy, conditional entropy and mutual information.

The information entropy H(*P*) of Knowledge *P* is defined by:

$$\mathrm{H}(P) = -\sum_{i=1}^{n} p(X_i)\log_2 p(X_i) \qquad (7)$$

The entropy is a nonnegative function, i, e, H(*P*) $\geq$ 0.It may be interpreted as a measure of the information content, or the uncertainty about knowledge *P* . Information entropy reaches a maximum value $\log|U|$,when the knowledge *P* becomes finest. The minimum value 0 is obtained, when the distribution of the knowledge *P* focuses on a particular value x_0, i, e. $p(x_0) = 1$ and $p(x) = 0, x \neq x_0$.

The conditional entropy H(*Q*|*P*) of the knowledge *Q* given by the knowledge *P* is expressed by:

$$\mathrm{H}(Q|P) = -\sum_{i=1}^{n} p(X_i)\sum p(Y_j|\mathrm{X_i})\log_2 p(Y_j|\mathrm{X_i}) \quad (8)$$

Conditional entropy is nonnegative and non-symmetric, namely, H(*Q*|*P*) $\geq$ 0 and in general H(*Q*|*P*) $\neq$ H(*P*|*Q*). It measures the additional amount of information provided by *Q* if *P* is known.

Mutual information can be defined by using entropy and conditional entropy as follows:

$$\mathrm{I}(P;Q)=\mathrm{H}(Q)-\mathrm{H}(Q|P). \qquad (9)$$

Mutual information measures the decrease of uncertainty about *Q* caused by *P*, and its inverse is the same. It measures the amount of information about *P* contained in *Q* or *Q* contained in *P*. The amount of information contained in *P* about itself is obviously H(*P*), namely, I(*P*;*Q*)= H(*P*).Attribute reduction depends on a criterion determining the attribute importance. By calculating the change in mutual information when an attribute is added to the set of considered condition attributes, a measure of the significance of the attribute can be obtained. The higher the change in mutual information is, the more significant the attribute is.

Theorem 1 [5] .Suppose DT=（*U*, *C*∪*D*, A, f ） is a decision table, where *U* is the universe of discourse ,*C* is the set of conditional attributes and *D* is the set of decision attributes.

For arbitrary set *B* $\subseteq$ *C*, the sufficient and necessary conditions that *B* is a relative reduction of *C* with respect to D are the followings, and the two conditions must be satisfied at the same time.

(1) I（*B*;*D*）=I（*C*;*D*）,

(2) H（*D*|*B*）< H（*D*|*B*-{*p*}）,for an arbitrary attribute *p* ∈ *B*.

2.2. *Discernibility Matrix*

Let *DT*=（*U*, *C*∪*D*, A, f ） be a decision table. By *M(DT)* we denote an $n \times n$ matrix(c_{ij}),called the discernibility matrix[6] of *DT*, such that:

$$c_{ij} = \begin{cases} \{\alpha \mid \alpha \in \mathrm{C} \wedge (f_\alpha(x_i) \neq f_\alpha(x_j))\}, & f_D(x_i) \neq f_D(x_j), \\ \lambda, & f_D(x_i) \neq f_D(x_j) \wedge f_C(x_i) = f_C(x_j), \\ \emptyset, & f_D(x_i) = f_D(x_j)。 \end{cases} \qquad (10)$$

Since *M(DT)* are symmetric and $c_{ii} \square \succ$ for *i*=1,2,...,*n*, we represent *M(DT)* only by elements in the lower triangle of *M(DT)*, respectively, i, e.The c_{ij}'s with $1 \square \ j < i \square \ n$.

Using discernibility matrix, Skowron and Rauszer have proven several properties and constructed efficient algorithms related to information systems and decision tables, e.g. The set of all indispensable attributes in *C* is called the core of *DT*, denoted by $CORE_c(D)$. $CORE_c(D)$ can be characterized by *M*(*DT*) in the following way:

$$CORE_c(D) \square \{a \,|\, (a \square \mathrm{C}) \square (\hat{}\, c_{ij}, ((c_{ij} \square M_{n\times n}) \square (c_{ij} \square \{a\})))\} \ . \qquad (11)$$

3. A Heuristic Algorithm for Reduction of Knowledge

A reduct is a subset of condition attributes that is jointly sufficient and individually necessary for preserving the same information under consideration as provided by the entire set of attributes. This algorithm attempts to find a minimal reduct without exhaustively generating all possible subsets. It starts with relative core and adds one attribute that results in the highest increase in H(*D*|*B*∪{ c_i }) in turn, until ending condition is met. This method does not always generate a minimal reduct, but it does result in a close-to-minimal reduct, which is still useful in reducing data set dimensionality.

Algorithm:
Input:
Decision table *DT*=（*U*, *C*∪*D*, A, f ）.
Output:
B , which is one relative reduct of conditional attribute set *C* with respect to decision attribute set *D*.
Step1:
Compute the mutual information I(*C*;*D*) between conditional attribute set *C* and decision attribute set *D* in the decision table *DT*;
Step2:
Compute the relative core of *C* with respect to *D* by discernibility matrix denoted by CORE $_D$ (*C*);
Step3:
B←CORE $_D$ (*C*);
Step4:
Compute I(*B*;*D*), if I(*B*;*D*)=I(*C*;*D*),go to step6, otherwise go to step5;
Step5:
$\square\, c_i \square$ *C*\B,compute the significance of attribute c_i ,and c_m =arg $\min\limits_{c_i \in C \backslash B}$ H(*D*|*B*∪{ c_i })(If multiple attributes achieving the maximum exist at the same time, choose one whose combination with B reaches the least as c_m),let *B*=*B*∪{ c_i },go to step4;
Step6:

Conditional attribute set B is a relative reduct we need.

4. An Illustrative Example

To illustrate the operation of attribute reduction, an example is given here (see Table I).

TABLE I. A DECISION TABLE

U	Conditional attribute set C				Decision Attribute D d
	Outlook (a1)	*Temperature (a2)*	*Humidity (a3)*	*Windy (a4)*	
1	Sunny	Hot	High	False	N
2	Sunny	Hot	High	True	N
3	Overcast	Hot	High	False	P
4	Rain	Mild	High	False	P
5	Rain	Cool	Normal	False	P
6	Rain	Cool	Normal	True	N
7	Overcast	Cool	Normal	True	P
8	Sunny	Mild	High	False	N
9	Sunny	Cool	Normal	False	P
10	Rain	Mild	Normal	False	P
11	Sunny	Mild	Normal	True	P
12	Overcast	Mild	High	True	P
13	Overcast	Hot	Normal	False	P
14	Rain	Mild	High	True	N

In the decision table defined in table1,U is the 14 objects, decision attributes set D is $\{d\}$,conditional attributes set is $\{ a_1, a_2, a_3, a_4 \}$and IND(D) ={{1,2,6,8,14},{3,4,5,7,9,10,11,12,13}}.

Step1: Calculate I(C;D)=H(D)- H（D|C）=0.940-0=0.940;

Step2: Calculate the discernibility matrix.

$$M_{14\times14}(DT) \square \begin{bmatrix} \square & & & & \\ \square & \square & & & \\ \vdots & \vdots & \ddots & & \\ \{a_1,a_3\} & \{a_1,a_3,a_4\} & \cdots & \square & \\ \square & \square & \cdots & \{a_1,a_2,a_3,a_4\} & \square \end{bmatrix}_{14\times14}.$$

And in the matrix:

$c_{1,1} \square \square ;$

$c_{2,1} \square c_{2,2} \square \square ;$

$c_{3,1} \square \{ a_1 \}, c_{3,2} \square \{ a_1, a_4 \}, c_{3,3} \square \square ;$

$c_{4,1} \square \{ a_1, a_2 \}, c_{4,2} \square \{ a_1, a_2, a_4 \}, c_{4,3} \square c_{4,4} \square \square ;$

…

$c_{14,1} \square c_{14,2} \square \square , c_{14,3} \square \{ a_1, a_2, a_4 \}, c_{14,4} \square \{ a_4 \}, c_{14,5} \square \{ a_2, a_3, a_4 \}, c_{14,6} \square \square , c_{14,7} \square \{ a_1, a_2, a_3 \}$ $, c_{14,8} \square \square , c_{14,9} \square \{ a_1, a_2, a_3, a_4 \}, c_{14,10} \square \{ a_3, a_4 \},$

$c_{14,11}$ □ $\{a_1, a_3\}, c_{14,12}$ □ $\{a_1\}, c_{14,13}$ □ $\{a_1, a_2, a_3, a_4\}$

$c_{14,14}$ □ □ ;

From the matrix, obviously $c_{3,1}$ □ $\{a_1\}, c_{14,4}$ □ $\{a_4\}$ is the single attribute element , thus CORE$_D$ (C)= $\{a_1, a_4\}$;

Step3: Let B ={ a_1, a_4 };

Step4:Calculate I(B;D)=H(D)-H(D|B)=0.940-0.679=0.261, obviously I(B;D)≠I(C;D);

Step5: Calculate H(D|$B\cup\{a_2\}$)= H(D|$B\cup\{a_3\}$)= 0,since the number of the combination with B of a_2 and a_3 is same. So a_2 or a_3 is chosen to be put into B,B=B ∪ { a_2 }={ a_1 , a_2 , a_4 }or B1=B ∪ { a_2 }={ a_1 , a_3 , a_4 } ,then go to step4 and calculate I(B;D)=H(D)- H(D|B) =0.940-0=0.940=I(C;D),I(B1;D)=H(D)- H(D|B1) =0.940-0=0.940= I(C;D),so it is the end. The result is $\{a_1, a_2, a_4\}$ and $\{a_1, a_3, a_4\}$.

5. Conclusions

In this paper, we address attribute reduction of rough set theory under the information-theoretic frame and discernibility matrix. Mutual information is presented. Based on the measure, an approach of attribute reduction based on rough sets is proposed. By constructing an example, we show how the technique works.

References

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

[2] PawlakZ.Rough set approach to multi-attribute decision analysis [J].European Journal of Operational Research,1994,72:443-459.

[3] Wenxiu Zhang,Weizhi Wu,Jiye Liang.Theory and Approach in Rough Set[M].Beijing:Science Press,2001(in Chinese).

[4] D.Q. Miao, J. Wang, An information representation of the concepts and operations in rough set theory, Journal of Software 10 (1999) 113-116 (in Chinese).

[5] D.Q. Miao,Daoguo Li.Theory,algorithm and application in rough set[M]. beijing:tsinghua university press.2008,4(in Chinese).

[6] Skowron, A, Rauszer, C., “The discernibility matrices and functions in information systems”, Slowifiski(Ed.), Intelligent decision support: Handbook of applications and advances of rough set theory, Kluwer Academic Publishers, Dordrecht, volume 11, 1992, pp.331-362.