# RESULT COMPARISON OF TWO ROUGH SET BASED DISCRETIZATION ALGORITHMS

Shanchan Wu, Wenyuan Wang
*Department of Automation, Tsinghua University,Beijing 100084, P.R.C.*

Keywords: Rought set, Cuts, Discretization, Data mining

Abstract: The area of knowledge discovery and data mining is growing rapidly. A large number of methods are employed to mine knowledge. Many of the methods rely of discrete data. However, most of the datasets used in real application have attributes with continuous values. To make the data mining techniques useful for such datasets, discretization is performed as a preprocessing step of the data mining. In this paper, we discuss rough set based discretization. We use UCI data sets to do experiments to compare the quality of Local discretization and Global discretization based on rough set. Our experiments show that Global discretization and Local discretization are dataset sensitive. Neither of them is always better than the other, though in some cases Global discretization generates far better results than Local discretization.

## 1 INTRODUCTION

Rough Set theory is a tool to tackle fuzzy and uncertainty knowledge. It was put forward firstly by Z.Pawlak (Pawlak Z, 1982). In decades, rough set theory has been successfully implemented in Data mining, artificial intelligence and pattern recognition.

But rough set and many other methods used in data mining can't deal with continuous attributes and a very large proportion of real data sets include continuous variables. One solution to this problem is to partition numeric variables into a number of intervals and treat each interval as a category. This process is usually termed dicretization.

Several methods have been proposed to discretize data as a preprocessing step for the data mining process. Nguyen S. H. proposed the named discretization approach based on rough set methods and boolean reasoning (Nguyen, 1995, 1997). The main idea is to seek possibly minimum number of discrete intervals, and at the same time it should not weaken the indiscernibility. It has been proven that Optimal Discretization Problem is NP-complete (Nguyen, 1995). In this paper, we examine two discretization algorithms based on rough set, Local Discretization and Global Discretization (Hung Son Nguyen, 1996). We do experiments to compare the results of the two algorithms.

This paper is organized as follows. In Section 2, we describe discretization based on rough set. Then we explain determination of candidate cuts and calculating of discernibility of cuts in Section 3. In Section 4, we describe the local discretization algorithm and global discretization algorithm and in section 5 we show the experiment results. Finally Section 6 concludes this paper.

## 2 DESCRIPTION OF ROUGH SET BASED DISCRETIZATION

An information system is defined as follows:

$$S = (U, A, V_a, F_a), \qquad a \in A \qquad (1)$$

where $U = \{x_1, x_2, \cdots, x_n\}$ is a finite set of objects(n is the number of objects), A is a finite set of attributes, $V = \bigcup_{a \in A} V_a$, and $V_a$ is a domain of attribute a, $F_a : U \times A \to V_a$ is a total function such that $f(x_i, a) \in V_a$ for each $a \in A$, $x_i \in U$.

An information system S in definition (1) is called a decision system or decision table when the attributes in S can be divided into condition attributes C and decision attributes D. i.e. $A = C \bigcup D$, and $C \bigcap D = \phi$.

In information systems, each subset of attributes $I \subseteq A$ determines a binary relation as follows:

$$IND(I) = \{< x, y > \in U \times U \mid \forall a \in I, a(x) = a(y)\}$$



Wu S. and Wang W. (2004).
RESULT COMPARISON OF TWO ROUGH SET BASED DISCRETIZATION ALGORITHMS.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 511-514
DOI: 10.5220/0002611505110514

It is easily shown that $IND(I)$ is an equivalence relation on the sets U and is called an indiscernible relation. The partition of U as defined by B will be denoted U/B and the equivalence classes introduced by B will be denoted $[u]_B$. In particular, $[u]_{\{d\}}$ will be called the decision classes of the decision system.

Let $\mathbf{S} = (U, A \cup \{d\}, V, f)$ be a decision table where $\mathrm{U} = \{\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \cdots, x_n\}$. Assuming that $\mathrm{V_a} = [I_a, r_a) \subset R$ for any $\mathrm{a} \in \mathrm{A}$ where R is the set of real numbers.

Assume now that the **S** is a consistent decision table. Let $\mathbf{D_a}$ be a partition of $\mathrm{V_a}$ (for $\mathrm{a} \in \mathrm{A}$) into subintervals, i.e.

$\mathrm{D_a} = \{[p_a^0 = l_a, p_a^1), [p_a^1, p_a^2), \cdots, [p_a^k, p_a^{k+1} = r_a)\}$, where

$\mathrm{V_a} = [p_a^0 = l_a, p_a^1) \cup [p_a^1, p_a^2) \cup \cdots \cup [p_a^k, p_a^{k+1} = r_a)$, and

$$p_a^0 = r_a < p_a^1 < p_a^2 < \cdots < p_a^k < p_a^{k+1} = r_a$$

Any $\mathrm{D_a}$ is uniquely defined by the set of cuts on $V_a : \{p_a^1, p_a^2 \cdots p_a^k\}$ (empty if card($\mathrm{D_a}$) = 1). The set of cuts on $\mathrm{V_a}$ defined by $\mathrm{D_a}$ can be identified by $\mathrm{D_a}$. A family $D = \{D_a : a \in V_a\}$ of partitions on S can be represented in the form $\bigcup_{a \in A} \{a\} \times D_a$

Any $(\mathrm{a}, \mathrm{v}) \in \mathrm{D}_a$ will be also called a cut on $\mathrm{V_a}$. Then the family $D = \{D_a : a \in V_a\}$ defines from $S = (U, A \cup \{d\})$ a new decision table $S^p = (U, A^p \cup \{d\})$, where $\mathrm{A^p} = \{a^p : a \in A\}$ and $\mathrm{a^p}(x) = i \Leftrightarrow a(x) \in [p_a^i, p_a^{i+1})$ for any $\mathrm{x} \in \mathrm{U}$ and $i \in \{0, \cdots k\}$.

After discretization, the original decision system is replaced with the new one. And different sets of cuts will construct different new decision systems. It is obvious that discretization process is associated with loss of information. Usually, the task of discretization is to determine a minimal set of cuts from a given decision system and keeping the discernibility between objects and the rationality of the selected cuts can be evaluated by the following criteria(Nguyen H S, 1995, 1997): (1) Consistency of P. For any objects $\mathrm{u}, \mathrm{v} \in \mathrm{U}$, they are satisfying if u, v are discerned by A, then u, v are discerned by P;(2) Irreducibility. There is no $\mathrm{P}' \subset \mathrm{P}$, satisfying the consistency; (3) Optimality. For any $\mathrm{P}'$ satisfying consistency, it follows $\mathrm{card(P)} \le card(\mathrm{P}')$, then P is optimal cuts. It has been proven that Optimal Discretization Problem is NP-complete (Nguyen H S, 1995).

# 3 DETERMINATION OF CANDIDATE CUTS AND CALCULATION OF DISCERNIBILITY OF CUTS

Let $\mathrm{S} = (U, A \cup \{d\}, V, f)$ be a decision system. An arbitrary condition attribute $\mathrm{a} \in \mathrm{A}$, defines a sequence $\mathrm{v_1^a} < \mathrm{v_2^a} < \cdots < \mathrm{v_{n_a}^a}$, where $\{\mathrm{v_1^a}, \mathrm{v_2^a}, \cdots, \mathrm{v_{n_a}^a}\} = \{a(x) : x \in U\}$, Then the set of all possible cuts on a is defined by:

$$\mathrm{C_a} = \left\{(\mathrm{a}, \frac{\mathrm{v_1^a} + \mathrm{v_2^a}}{2}), (\mathrm{a}, \frac{\mathrm{v_2^a} + \mathrm{v_3^a}}{2}), \cdots, (\mathrm{a}, \frac{\mathrm{v_{n_a-1}^a} + \mathrm{v_{n_a}^a}}{2})\right\}.$$

The set of all possible cuts on all attributes is denoted by: $\mathrm{C_A} = \bigcup_{a \in A} C_a$ .This method usually generates a large set of candidate cuts. In order to reduce the number of candidate cuts, we can use bound cuts (Jian-Hua Dai, 2002).

Since we are only interested in separating objects that have different decision values, each cut in our representation is given information about how many objects from each decision class are to the left and to the right of the cut, i.e. how many pairs of objects with different decision values that are discerned from each other. The algorithm where this measure is later used sequentially deals with each attribute and the set of cuts that may be introduced on that attribute. By assuming that we can totally order all objects so that they primarily are sorted on the value of the current attribute and secondly in some arbitrary order, we use the algorithm 1 to calculate

Table 1: Discernibility Conventions the discernibility value of a cut.

| |
|---|
| **(a,c)**：A cut point c on an attribute **a** dividing all objects in a decision system in two parts $\{u \in U : a(u) < c\}$ and $\{u \in U : a(u) > c\}$ |
| **D**： A set of cuts (a,c) |
| **AllCuts**：All possible cuts on the decision system |
| **L**： $U/\{(a,c) \in D\}$ or $U/B$, $B \subseteq A$ |
| $l^X(a,c), r^X(a,c)$ ： number of elements that are to the left/right of the cut (a, c) in the equivalence class X |
| $l_i^X(a,c), r_i^X(a,c)$ ：number of elements with decision value i in equivalence class X that are to the left/right of the cut (a, c) |
| $c_j$ ：a value indicating where the cut is made |

Before turning to the details of algorithm 1, internal notations used in the algorithm are explained in table 1. These notations are also used in the next section.

**Algorithm 1: Calculate the discernibility value of a cut**

Input: information system $\mathbf{S} = (U, A \cup \{d\})$, candidate cut $(a, c_j)$ on condition attribute $a \in A$

Output: the discernibility value of the cut $(a, c_j)$

Method:

$N \leftarrow 0$

**for each** $X \in L$ **do**

$$N \leftarrow N + \left(l^X(a,c) \cdot r^X(a,c)\right) - \sum_{i=1}^{|d|} l_i^X(a,c) \cdot r_i^X(a,c)$$

**return** N

The discernibility value returned by algorithm 1 is equal to the number of pairs of objects from S discerned by cut $(a,c_j)$. The proof that this algorithm is correct can be found in (Hung Son Nguyen, 1996).

# 4 LOCAL DISCRETIZATION AND GLOBAL DISCRETIZATION

In this section, we describe Local Discretization and Global Discretization algorithms (Hung Son Nguyen, 1996). Local Discretization algorithm works by finding a maximally discerning cut (see algorithm 1) from the set of all possible cuts (AllCuts) and then dividing the dataset into two subsets as long as there are objects with different decision values.

**Algorithm 2**. Local discretization

Input: information system $\mathbf{S} = (U, A \cup \{d\})$, all candidate cuts in **S**

Output: new information system after being discretized

Method:

$N_{UM}C_{LASSES}(S)$

Return number of decision classes in S

$T_{RAVERSE}(S)$

If $N_{UM}C_{LASSES}(S) > 1$ Then

from AllCuts select cut point $(a^*, c^*)$

which has maximal discernibility value using algorithm 1

$D \leftarrow D \cup \{(a^*, c^*)\}$

$AllCuts \leftarrow AllCuts \setminus \{(a^*, c^*)\}$

$U_1 \leftarrow \{x \in U : a^*(x) < c\}$

$U_2 \leftarrow \{x \in U : a^*(x) \geq c\}$

$T_{RAVERSE}(U_1)$

$T_{RAVERSE}(U_2)$

$L_{OCAL}D_{ISCRETIZATION}(S)$

AllCuts $\leftarrow$ the set of all possible cuts in S

$D \leftarrow \phi$

$T_{RAVERSE}(U)$

Discretize S using the cuts in D

In algorithm 3 (Hung Son Nguyen, 1996), it works with decision classes and check each consecutive cut that is added to the final set D against all objects that are not completely separated into equivalence classes uniform w.r.t. decision value by the current set of cuts. It splits the decision classes into smaller and smaller parts until they are uniform with respect to the decision values of the objects.

**Algorithm 3**. Global Discretization

Input: information system $\mathbf{S} = (U, A \cup \{d\})$, all candidate cuts in **S**

Output: new information system after being discretized

Method:

$N_{UM}C_{LASSES}(S)$

Return number of decision classes in S

$G_{LOBAL}D_{ISCRETIZATION}(S)$

AllCuts $\leftarrow$ the set of all possible cuts in S

$D \leftarrow \phi$

$L \leftarrow U / B$, B is the set of attributes that will not be discretized

**repeat**

from AllCuts select cut point $(a^*, c^*)$

which has maximal discernibility value using algorithm 1

$D \leftarrow D \cup \{(a^*, c^*)\}$

$AllCuts \leftarrow AllCuts \setminus \{(a^*, c^*)\}$

**for each** $X \in L$ **do**

$L \leftarrow L \setminus \{X\}$

if $N_{UM}C_{LASSES}(X) > 1$ then

$X_1 \leftarrow \{x \in X : a^*(x) \leq c^*\}$

$X_2 \leftarrow \{x \in X : a^*(x) > c^*\}$

$L \leftarrow L \cup \{X_1, X_2\}$

**until** $L = \phi$

# 5 EXPERIMENTS AND ANALYSIS

We do our experiments on three data sets from UCI named abalone and iris and liver disorders respectively, which can be downloaded from the website (MLR). Some information about the data sets is shown in table 2.

Table 2: The data sets for experiments

| Name | #objects | #continuous attributes | #decision classes |
|---|---|---|---|
| Iris | 150 | 4 | 3 |
| liver-disorders | 345 | 6 | 2 |
| Abalone | 4177 | 7 | 29 |

We make comparative experiments between local discretization algorithm and global algorithm, comparing the number of result cuts discretizing continuous attributes. The results are shown in table 3, table 4, and table 5 respectively. In the tables, #cuts L denotes the number of result cuts generated by local discretization algorithm and #cuts G by global discretization algorithm.

As the two algorithms are both applied on consistent information systems and maintain the original indiscernibility, the smaller number of the result cuts, the better the algorithm is. From the comparisons we know that for liver disorders dataset and abalone dataset, the number of result cuts generated by global algorithm is far smaller than by local algorithm. But it is larger for liver iris dataset. So we can't say that global algorithm is always better than local algorithm.

For liver iris data set, the number of result cuts of attribute sepal_length generated by global algorithm is far larger than by local algorithm, and the number of result cuts of other attributes is almost equal. But for two other data sets, the number of result cuts for all attributes generated by global algorithm is far smaller than by local algorithm. Hence, we can say that the two algorithms are data set sensitive, and we can conjecture that their quality depends on the distributions of the values of the attributes and their decision classes.

Table 3: Comparison of the results on liver disorders.

| Attribute | Mcv | alkphos | sgpt | Sgo | Gam-magt | drinks | total |
|---|---|---|---|---|---|---|---|
| #cuts L | 20 | 22 | 20 | 25 | 30 | 23 | 140 |
| #cuts G | 3 | 4 | 3 | 2 | 5 | 3 | 20 |

Table 4: Comparison of the results on iris.

| Attribute | sepal_ Length | sepal_ width | petal_ length | petal_ width | Total |
|---|---|---|---|---|---|
| #cuts L | 3 | 3 | 6 | 1 | 13 |
| #cuts G | 34 | 2 | 4 | 2 | 42 |

Table 5: Comparison of the results on abalone

| Attri-bute | len-gth | diam-eter | hei-ght | Whole weight | shucked weight | viscera weight | shell weight | total |
|---|---|---|---|---|---|---|---|---|
| #cuts L | 421 | 389 | 419 | 539 | 564 | 674 | 555 | 3561 |
| #cuts G | 20 | 21 | 30 | 7 | 32 | 32 | 30 | 172 |

## 6 CONCLUSIONS

For discretization based on rough set, we should seek possible minimum number of discrete internals, and at the same time it should not weaken the indiscernibility ability. This paper examines two algorithms (Hung Son Nguyen,1996), local discretization and global discretization. Our experiments show that the discretization algorithms are dataset sensitive. Neither of them always generates smaller number of result cuts. On some datasets, one algorithm generates fewer result cuts, but on other datasets it is contrary. We can conjecture that the quality of the two algorithms depends on the distributions of the values of the continuous dataset attributes and their decision classes. How the distributions affect the results is what we will study further. With that, we can use some methods to improve the algorithms.

## REFERENCES

Pawlak Z (1982, November 5). Rough Sets. Int'l J. Computer & Science [J], 11, 341-356.

Nguyen H S, Skowron A (1995). Quantization of real value attributes. Proceedings of Second Joint Annual Conf. on Information Science, Wrightsville Beach, North Carolina, 34-37.

Nguyen H S (1997). Discretization of Real Value Attributes: Boolean reasoning Approach [PhD Dissertation]. Warsaw University Warsaw, Poland.

Hung Son Nguyen, Sinh Hoa Nguyen (1996). Some efficient algorithms for rough set methods. In 6th International conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, 1451-1456.

Jian-Hua Dai, Yuan-Xiang Li (2002, November 4-5). Study on discretization based on rough set theory. Proceedings of the First International Conference on Machine Learning and Cybernetics, 3, 1371-1373.

MLR, http://www.ics.uci.edu/~mlearn/MLRepository.html.