IEMS Vol. 10, No. 1, pp. 65-73, March 2011.

Rule Induction Considering Implication Relations Between Conclusions

Masahiro Inuiguchi†
Department of Systems Innovation
Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama-cho, Toyonaka, Osaka 563-0057, JAPAN
Tel: +81-6-6850-6350, E-mail: inuiguti@sys.es.osaka-u.ac.jp

Masanori Inoue
Department of Systems Innovation
Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama-cho, Toyonaka, Osaka 563-0057, JAPAN
Tel: +81-6-6850-6350

Yoshifumi Kusunoki
Division of Electrical, Electronic and Information Engineering
Graduate School of Engineering, Osaka University
2-1 Yamadaoka, Suita, Osaka, 565-0871, Japan
Tel: +81-6-6879-7787, E-mail: kusunoki@eei.eng.osaka-u.ac.jp

Received, January 18, 2011; Revised, February 17, 2011; Accepted, February 21, 2011

Abstract. In rough set literatures, methods for inducing minimal rules from a given decision table have been proposed. When the decision attribute is ordinal, inducing rules about upward and downward unions of decision classes is advantageous in the simplicity of obtained rules. However, because of independent applications of the rule induction method, inclusion relations among upward/downward unions in conclusion parts are not inherited to the condition parts of obtained rules. This non-inheritance may debase the quality of obtained rules. To ensure that inclusion relations among conclusions are inherited to conditions, we propose two rule induction approaches. The performances of the proposed approaches considering the inclusion relations between conclusions are examined by numerical experiments.

Keywords: Rough Set, Rule Induction, Upward/Downward Union, MLEM2

1. INTRODUCTION

From a given data table, we may induce rules by various methods (for example, see Wikipedia Rule Induction). Among those rule induction methods, we focus on rough set based rule induction (Pawlak, 1991). In the rough set based rule induction, methods for inducing minimal rules from a given decision table have been pro-posed. A decision table is a collection of objects expressed by combinations of profiles and decision classes. The profiles are composed of conditions on explanatory variables called condition attributes. Some data are inconsistent one another so that two or more objects having the same profiles are classified into different decision classes. Such inconsistencies are processed by approximation operations so that the lower approximation of a decision class includes only objects consistent with others while the upper approximation of a decision class includes its original members and their conflicting members. Those approximations are consistent in the sense that there is no conflict with their complements.

Using the approximations, rough set rule induction can be seen as a procedure to find the minimal conditions separating members of an approximation from non-members. As such algorithms, LEM2-based algorithms were proposed under many situations. These algorithms do not induce all decision rules underlying in

† : Corresponding Author

the given decision table but induce a minimal set of rules by which all members are explained. The original LEM2 proposed by Grzymala-Busse (1992) induces rules from a given decision table when all attributes are nominal. Then Grzymala-Busse (2003) proposed MLEM2 in order to treat a case when some of condition attributes are ordinal (numerical). Stefanowski (1988) previously proposed MODLEM for the same purpose. Greco *et al*. (2002) proposed DOMLEM in order to treat a case when the monotonicity between ordinal condition attributes and an ordinal decision attribute is supposed. The dominance-based rough set approach (DRSA) (Greco *et al*., 1998; Greco *et al*., 1999; Greco *et al*., 2001) is used for the underlying rough set model of DOMLEM. In this paper, MLEM2 is used as rule induction algorithm but the proposed idea can be applied to any rule induction algorithms.

We consider a case when some of condition attributes and the decision attribute are ordinal. Because of the ordinal property of the decision attribute, it is advantageous in the sense of simplicity of the obtained rules to induce rules with respect to upward and downward unions of decision classes than to induce rules with respect to decision classes directly. Following the conventional approach, we may apply MLEM2 to induce rules with respect to upward and downward unions of decision classes, independently.

We note that, when conclusions of rules include upward or downward unions of decision classes, two of them may have the inclusion relations. For example, considering rules for "not less than 2" and rules for "not less than 1", conclusion "not less than 1" includes conclusion "not less than 2". When decision rules are induced independently, it may happen that the inclusion relations are not inherited in the premises. This non-inheritance may debase the quality of obtained rules.

Considering this non-inheritance, in this paper, we proposed two methods to inherit the inclusion relations among the conclusions to the premises of decision rules. One is the refining approach and the other is the coarsening approach. Because we may apply different approaches to upward and downward unions, we consider four combinations. By numerical experiments with datasets showing hill structures, these four combinations and the conventional approach are compared in the classification accuracy of the induced set of rules. Moreover, we further extend the proposed approach to datasets showing mixed-case of hill and valley structures. We would observe that the same results when the standard decision attribute value is known.

This paper is organized as follows. In section 2, rough set approaches to rule induction are briefly reviewed. The proposed approaches considering the implication relations between conclusions of rules are explained in section 3. In section 4, the results of numerical experiments are described. Concluding remarks are given in section 5.

2. ROUGH SET BASED RULE INDUCTION

2.1 Decision Tables

In the rough set theory, decision tables showing object features are analyzed. A decision table is formally characterized by a quadruple $S = \langle U, C \cup \{d\}, V, f \rangle$, where U is a finite set of objects, C is a finite set of condition attributes, $d \notin C$ is a unique decision attribute, $V = \bigcup_{a \in C \cup \{d\}} V_a$ is a set of attribute values, V_a is a set of values with respect to attribute a and f: $U \times C \cup \{d\} \rightarrow V$ is a total function called an information function. Objects in a decision table are classified by their decision attribute values into decision classes Cl_i, $i = 1, 2, \cdots, p$. We assume that each attribute is nominal or ordinal. Values of a nominal attribute are considered labels and used to distinguish objects. Therefore, there is neither order relation between nominal attribute values nor the magnitude of the difference. On the other hand, values of an ordinal attribute are assumed to be totally ordered, however a magnitude of their difference is meaningless.

2.2 Decision Rules and Rough Sets

A decision rule is an if-then rule composed of a premise and a conclusion. It is described as "**IF** *premise* **THEN** *conclusion.*" The premise of a decision rule is expressed as a conjunction of elementary conditions. The elementary condition is represented as '$f(a, x) = v$' ($v \in V_a$) for a nominal condition attribute a while it is represented as '$f(a, x) \geq v^{\mathrm{L}}$' or '$f(a, x) < v^{\mathrm{R}}$' ($v^{\mathrm{L}}, v^{\mathrm{R}} \in V_a$) for an ordinal/numerical condition attribute a. The conclusion is represented as '$x \in Cl_i$' or '$x \in \bigcup_i Cl_i$' in this paper. Decision rules are induced based on objects in the given table by generalizing their conditions. A decision rule is simply called a rule if there is no confusion.

When a given decision table includes at least two objects which share the same condition attribute values but take different decision attribute values, those objects are inconsistent. In this case, we cannot induce exact rules corresponding to those objects. Then we apply the rough set theory (Pawlak, 1982; Pawlak, 1991). In the rough set theory, the inconsistency is treated reasonably by lower and upper approximations of decision classes.

Now let us introduce lower and upper approximations. First, we define an equivalence class of an object x with respect to C by

$$[x]_C = \{y \in U : f(y, a) = f(x, a), \forall a \in C\}. \quad (1)$$

Then lower and upper approximations of Cl_i are defined by

$$C_*(Cl_i) = \{x \in U : [x]_C \subseteq Cl_i\}, \quad (2)$$

$$C^*(Cl_i) = \{x \in U : [x]_C \cap Cl_i \neq \phi\}. \quad (3)$$

The pair ($C_*(Cl_i)$, $C^*(Cl_i)$) is called a rough set of Cl_i.

The lower approximation of a decision class includes only consistent objects while the upper approximation includes possible objects. Then objects in lower approximations as well as those in upper approximations are consistent in the sense that there is no conflict between members and nonmembers.

When the given decision table includes inconsistent objects, we induce rules with respect to Cl_i. (resp. $\bigcup_i Cl_i$) based on objects in its lower approximation $C_*(Cl_i)$ (resp. $C_*(\bigcup_i Cl_i)$) or those in its upper approximation $C^*(Cl_i)$ (resp. $C^*(\bigcup_i Cl_i)$). Rules induced based on lower approximations are called certain rules while rules induced based on upper approximations are called possible rules. In this paper, we focus on certain rules.

2.3 Rule Induction Algorithm

In this paper, we study methods for inducing rules whose premises reflect the implication relations between conclusions. We use MLEM2 proposed by Grzymala-Busse (2003) as the fundamental rule induction algorithm. The algorithm of MLEM2 is shown in Figure 1. First, a target decision class or a set of decision classes is approximated by the manner of rough set theory. In this algorithm, B is the input and it is defined by a lower approximation or an upper approximation of a certain class or a union of classes. The algorithm outputs ***T***, a minimal set of minimal rules. The minimal set of decision rules means that there exists no redundant rule in itself. On the other hand, the minimal rule means there exists no redundant condition in its premise. The algorithm is based on the sequential covering method (Fürnkranz, 1999) equipped with the general to specific search (Hoover and Perez, 1999).

```
Input: a set B
Output: a single local covering T of B
 1: T := ∅;
 2: G := B;
 3: while G ≠ ∅ do
 4:    T := ∅; T(G) := {t | [t] ∩ G ≠ ∅};
 5:    while T = ∅ or [T] ⊈ B do
 6:       † select t ∈ T(G) with the highest priority, if a
          tie occurs, select t ∈ T(G) such that |[t] ∩ G| is
          maximum; if another tie occurs, select t ∈ T(G)
          with smallest cardinality of [t]; if a further tie occurs,
          select a first one;
 7:       T := T ∪ {t};
 8:       G := [t] ∩ G;
 9:       T(G) := {t | [t] ∩ G ≠ ∅};
10:       T(G) := T(G) − T;
11:    end while
12:    for each t in T do
13:       if [T − {t}] ⊆ B then
14:          T := T − {t};
15:       end if
16:    end for
17:    T := T ∪ {T};
18:    G := B − ⋃_{T∈T}[T];
19: end while
20: for each T in T do
21:    if ⋃_{S∈T−{T}}[S] = B then
22:       T := T − {T};
23:    end if
24: end for
```

Figure 1. Algorithm of MLEM2.

The algorithm iteratively adds the best condition in a given greedy criterion until objects satisfying all conditions belong to B, and removes those objects from target set G. Redundant conditions are removed at first screening process at lines 12 to 16. Taking a conjunction of remaining conditions, a rule having the conjunction as its premise is induced. This iteration written from line 3 to line 19 of Figure 1 repeats until G becomes empty. The greedy criterion is lexicographical as specified at line 6. From the first criterion, a condition satisfied with the most objects in target set G is selected. This can be seen as a general to specific search. The lines 20 to 24 are the second screening process which removes redundant rules.

2.4 Upward and Downward Unions

In the real world, the decision attribute is often ordinal. In evaluation problems, for example, we may rank objects into three classes, i.e., bad, medium and good. In this case, we presume an order bad $\prec$ medium $\prec$ good. Namely class 'bad' is worse than class 'medium,' class 'medium' is worse than class 'good' and the transitivity is naturally assumed.

When the decision attribute is ordinal, inducing decision rules with respect to upward and downward unions is more advantageous in the simplicity and applicability than inducing decision rules with respect to decision classes. Under an order $Cl_1 \prec Cl_2 \prec \cdots \prec Cl_p$, the upward union $Cl_t^{\geq}$ and $Cl_t^{\leq}$ are defined by

$$Cl_t^{\geq} = \bigcup_{i \geq t} Cl_i, \quad Cl_t^{\leq} = \bigcup_{i \leq t} Cl_i. \quad (4)$$

The induction of decision rules with respect to upward and downward unions can be done in the same way as that with respect to decision classes. However, because of independent applications of a rule induction algorithm such as MLEM2, the inclusion relations such as $Cl_t^{\geq} \subseteq Cl_s^{\geq}$ and $Cl_t^{\leq} \supseteq Cl_s^{\leq}$ for $t > s$ would not be reflected in the induced rules.

3. PROPOED APPROACHES

3.1 Refining Approach

Let X and Y be conclusions of rules and X imply Y.

This approach is based on the idea that the premises of rules with X should be stronger than the premises of rules with Y. To crystallize this idea, after inducing all rules having conclusion Y, rules having conclusion X are induced by refining the premises of rules having conclusion Y. In MLEM2, we may realize this approach by the following changes: (1) we replace "$T := \varnothing$"; by "$T := T'$; $G := [T] \cap G$;" at line 4 of Figure1, where T' is the premise of a rule with Y. (2) Before this command, we should select T' from premises of all rules with Y by the same criterion described at line 6. (3) Moreover we must modify the screening of T so that we maintain the relation that T implies T'.

3.2 Coarsening Approach

In contrast to the refining method, this method is based on the idea that the premises of rules with Y should be weaker than the premises of rules with X when X implies Y. To crystallize this idea, after inducing all rules having conclusion X, rules having conclusion Y are induced by coarsening the premises of rules having conclusion X. In MLEM2, we may realize this approach by the following changes: (1) we replace "$T(G) := \{t \mid [t] \cap G\}$" by "$T(G) := \{t \mid [t] \cap G, T' \text{ implies } t\}$" at lines 4 and 9 of Figure 1, where T' is the premise of a rule with X. (2) Moreover we must modify the screening of $\boldsymbol{T}$ so that we maintain the relation that for each rule with X, there exists at least one rule with Y whose premise is implied by the premise of the rule with X.

3.3 Combinations of Approaches for Upward and Downward Unions

In the previous subsection, we have described rule induction approaches to upward/downward unions. However, we have not yet mentioned about the rule induction approaches to obtain both rules with respect to upward unions and rules with respect to downward unions. Unless both rules with respect to upward unions and rules with respect to downward unions are induced, we would not infer the membership to decision class from condition attribute values of a new object with a sufficient accuracy. In this paper, we consider four combinations of three approaches described in the previous subsection. The four combinations are shown in Table 1.

Table 1. Four possible combinations

Comb.	Upward unions	Downward unions
CA1	Refining approach	Refining approach
CA2	Refining approach	Coarsening approach
CA3	Coarsening approach	Refining approach
CA4	Coarsening approach	Coarsening approach

As shown in Table 1, the first combined approach, CA1 uses the refining approach to both rule induction with respect to upward and downward unions. The second combined approach, CA2 uses the refining approach to rule induction with respect to upward unions and the coarsening approach to rule induction with respect to downward unions. The third combined approach, CA3 is the opposite to CA2 and uses the coarsening approach to rule induction with respect to upward unions and the refining approach to rule induction with respect to downward unions. The fourth combined approach, CA4 uses the coarsening approach to both rule induction with respect to upward and downward unions.

3.4 Application of Induced Rules to New Objects

Once we obtain rules with respect to upward unions as well as those with respect to downward unions, we can use those rules to classify a new object into a decision class. However, due to the lack of the comprehensiveness and total consistency of data in a given decision table, induced rules are often imperfect. They would have some conflicts, inapplicabilities and indecisiveness. Therefore, we propose a classification algorithm to resolve those difficulties so that we estimate a single class in which the new object may be included. The resolution algorithm consists of two steps.

First, for $t = 2, \cdots, p$ we decide whether the new object x is classified into $Cl_t^{\geq}$ or $Cl_{t-1}^{\leq}$ using evaluation measures $EM(Cl_t^{\geq})$ and $EM(Cl_{t-1}^{\leq})$, defined by

$$EM(Cl_t^{\geq}) = \sum_{s \geq t} Supp(Cl_s^{\geq}), \tag{5}$$

$$EM(Cl_{t-1}^{\leq}) = \sum_{s \leq t-1} Supp(Cl_s^{\leq}), \tag{6}$$

where $Supp(X)$ is a measure called support defined by

$$Supp(X) = \sum_{\substack{\text{matching rules} \\ r \text{ inferring } X}} Strength(r) \times Specificity(r) \tag{7}$$

$Strength(r)$ proposed in LERS (1992) is the total number of objects in the given decision table correctly classified by rule r while $Specificity(r)$ is the total number of conditioned attribute variables in the premise of rule r. If $EM(Cl_t^{\geq}) \geq EM(Cl_{t-1}^{\leq})$ then object x is classified to $Cl_t^{\geq}$; otherwise to $Cl_{t-1}^{\leq}$. However, it could happen that there is no rule inferring $\{ Cl_s^{\geq},\ s \geq t \}$ or $\{ Cl_s^{\leq},\ s < t \}$ and matching to x, namely, $EM(Cl_t^{\geq}) = EM(Cl_{t-1}^{\leq}) = 0$. In this case, we use auxiliary evaluation measures $EM'(Cl_t^{\geq})$ and $EM'(Cl_{t-1}^{\leq})$, defined by

$$EM'(Cl_t^{\geq}) = \sum_{s \geq t} M(Cl_s^{\geq}), \tag{8}$$

$$EM'(Cl_{t-1}^{\leq}) = \sum_{s \leq t-1} M(Cl_s^{\leq}), \tag{9}$$

where $M(X)$ is defined by

$$Supp(X) = \sum_{\substack{\text{partially matching} \\ \text{rules } r \text{ inferring } X}} Matching\_factor(r) \times Strength(r) \times Specificity(r) \quad (10)$$

And *Matching_factor*(*r*) is defined by the ratio of matched elementary conditions to all elementary conditions of rule *r*. If $EM'(Cl_t^{\geq}) \geq EM'(Cl_{t-1}^{\leq})$ then *x* is classified into $Cl_t^{\geq}$; otherwise $Cl_{t-1}^{\leq}$.

Second, we aggregate results in the first step. We adopt a majority voting. If *x* is classified into $Cl_t^{\geq}$ (and not into $Cl_{t-1}^{\leq}$) in the first step, then all of Cl_t, Cl_{t+1}, ⋯, Cl_p get a vote. On the other hand, if *x* is classified into $Cl_{t-1}^{\leq}$ in the first step, then all of Cl_1, Cl_2, ⋯, Cl_{t-1} get a vote. Then *x* is finally classified into a decision class receiving the highest number of votes. If a tie occurs, it is broken by a random selection among tied classes. For example, there exist four classes Cl_1, Cl_2, Cl_3 and Cl_4, and a new object *x* is classified into $Cl_2^{\geq}$, $Cl_3^{\geq}$, and $Cl_4^{\leq}$ in the first step, then the numbers of votes for Cl_1, Cl_2, Cl_3 and Cl_4 become 1, 2, 3 and 2, respectively. Consequently, *x* is classified into Cl_3.

4. NUMERICAL EXPERIMENTS

4.1 Outline

In order to examine the performances of the proposed approaches, we executed numerical experiments. We compare the proposed approaches with the conventional approach in MLEM2, i.e., independent applications of MLEM2 to each of upward and downward unions. We apply the proposed approaches as well as the conventional approach to a set of training data. We then evaluate the performance of the obtained set of rules by a set of checking data. The performance is measured by classification accuracy, i.e., the rate of objects correctly classified, to all checking data.

To execute the experiments, we need dataset suitable for proposed approaches. However, unfortunately, it is not very easy to obtain such dataset from the public domain, even though we often to come across them in the real world. Then we artificially generate datasets.

The experiments are made by two stages. In the first stage, we examined the proposed approaches in hill-structured datasets such that several mountains exist on the hyperplane. By the experiments of such datasets, we may find some inclinations of the proposed approaches. Then in the second stage, we examined the proposed approaches in hill-valley-structured datasets such that both mountains and valleys exist on the hyperplane. By the second stage experiments, we may confirm the findings of the first stage experiments.

4.2 Experiments by Hill-structured Dataset

4.2.1 Dataset Generation

To obtain the hill-structured dataset, we randomly generate hyper rectangles $H_1 \supseteq \cdots \supseteq H_p$, where *p* is a given number of decision class and H_t, $t = 1, \cdots, p$ are subsets of the Cartesian product of attribute value sets of condition attributes. We then regard a point in the Cartesian product as the profile (condition attribute values) of an object. The decision class Cl_t is composed of points in $H_t - H_{t+1}$, where, for convenience, we define $H_0 = \varnothing$. Several sequences of hyper rectangles $H_{j1} \supseteq \cdots \supseteq H_{jp}$, $j = 1, \cdots, q$ are independently generated. Then some hyper rectangles of different sequences may have intersections. The decision attribute value of an object *x* in an intersection is determined by the largest value *t* among H_{jt}'s such that $x \in H_{jt}$. Figure 2 shows an example of the nested structure when $|C| = 2$ and $p = 4$. In Figure 2, the decision attribute takes four values, ◎, ○, △ and × which are ordered as × ≺ △ ≺ ○ ≺ ◎.

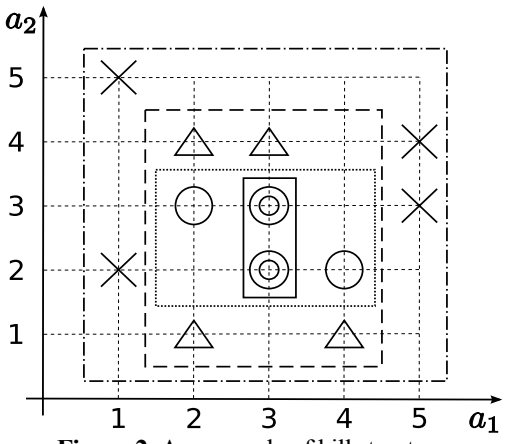

Figure 2. An example of hill structure.

We generate all points in H_{j1}, $j = 1, \cdots, q$ and collect pairs of condition attribute values and the decision attribute value of them and build a dataset. The set of training data are composed of sampled data from the dataset and the checking data are composed of the remaining data.

4.2.2 Results and Discussion

Results in four decision tables (datasets) which vary in numbers of condition attributes and numbers of its values are shown in Table 2~Table 5. All decision tables have four ordered decision classes and three sequences of hyper rectangles. All condition attributes in each decision table are ordered. We executed the experiments with several different sampling rates for training data. The sampling rates are 1%, 2%, ⋯, 8%, 9%. In each table, row 'Rate' shows the sampling rates and row 'ML' indicates the results of the conventional approach based on MLEM2, i.e., the independent applications to upward/downward unions. Rows 'CA1,' 'CA2,' 'CA3' and 'CA4' show results of the proposed approaches of Table 1.

Table 2. Classification accuracy when $|C| = 5$, $|V_a| = 4$, $p = 4$ and 3 sequences of hyper rectangle.

Rate	1%	2%	3%	4%	5%	6%	7%	8%	9%
ML	34.2±7.3	43.3±7.7	51.2±6.4	55.6±7.5	61.0±7.2	65.7±5.8	67.1±6.1	70.2±6.1	73.1±6.4
CA1	32.9±7.2	42.3*±7.2	50.5*±7.3	54.0±7.5	59.1±6.6	63.1±6.1	65.6±6.7	68.6±6.2	71.7±6.3
CA2	34.0*±8.0	43.8*±8.1	53.6±7.5	56.6*±7.4	61.7*±6.3	66.4*±6.1	68.2±6.5	70.4*±5.8	73.5*±5.9
CA3	32.4±6.6	41.3±7.0	47.5±6.9	51.4±6.8	56.2±6.9	61.4±6.6	64.1±6.9	67.3±6.8	69.3±6.9
CA4	33.5*±6.9	41.4±6.5	48.9±7.0	52.6±6.5	57.9±6.8	62.4±6.8	65.6±7.1	68.3±6.8	70.3±6.9

Table 3. Classification accuracy when $|C| = 5$, $|V_a| = 5$, $p = 4$ and 3 sequences of hyper rectangle.

Rate	1%	2%	3%	4%	5%	6%	7%	8%	9%
ML	48.1±8.3	64.1±8.1	76.7±6.5	83.2±5.8	88.2±4.9	92.7±4.3	95.6±3.0	96.6±2.6	97.3±2.5
CA1	47.7*±9.0	64.1*±10.0	76.3±8.4	83.3*±6.8	89.4±5.2	93.1*±4.5	95.9*±3.4	97.0*±2.5	97.9±2.4
CA2	49.4*±9.7	67.5±9.0	80.5±8.3	86.9±6.4	92.1±5.5	96.0±3.9	97.8±2.9	98.6±2.1	98.9±1.8
CA3	46.7±9.5	61.5±10.0	73.8±8.5	81.4±7.1	87.6*±5.7	91.9±5.1	95.2*±3.6	96.6*±3.1	97.6*±2.9
CA4	47.2*±9.5	64.5±10.2	77.3*±7.8	84.6±7.0	90.2±5.5	94.3±4.6	96.9±3.5	98.1±2.6	98.6±2.2

Table 4. Classification accuracy when $|C| = 6$, $|V_a| = 4$, $p = 4$ and 3 sequences of hyper rectangle.

Rate	1%	2%	3%	4%	5%	6%	7%	8%	9%
ML	47.1±6.8	65.2±6.9	76.6±5.9	85.7±4.9	90.6±4.0	93.6±4.0	96.0±2.6	97.9±2.1	98.4±1.7
CA1	46.9*±7.3	62.9±7.6	74.3±7.5	83.9±5.3	88.8±4.8	92.0±3.7	94.6±2.8	96.6±2.8	97.7±2.2
CA2	51.2±8.8	68.4±8.2	81.6±7.3	90.5±5.7	94.1±4.1	96.7±3.1	98.1±2.1	99.2±1.2	99.5±0.9
CA3	43.7±6.9	58.6±7.6	69.7±6.7	81.2±6.6	86.8±5.7	90.8±5.1	93.7±3.4	96.4±3.1	97.3±2.4
CA4	46.8*±7.9	62.7±8.1	75.7*±7.0	86.4*±6.9	91.6±5.2	94.9±4.5	97.1±3.3	98.9±1.7	99.1±1.5

Table 5. Classification accuracy when $|C| = 6$, $|V_a| = 5$, $p = 4$ and 3 sequences of hyper rectangle.

Rate	1%	2%	3%	4%	5%	6%	7%	8%	9%
ML	72.1±4.9	86.8±3.6	92.7±2.3	95.5±1.6	96.9±1.2	97.8±0.9	98.6±0.7	98.8±0.6	99.1±0.6
CA1	70.7±5.8	85.5±3.7	91.8±2.7	94.9±1.8	96.6±1.5	97.6*±1.2	98.3±0.9	98.6±0.8	98.9±0.7
CA2	74.5±6.2	88.1±3.2	93.8±2.2	96.1±1.6	97.5±1.3	98.3±1.0	99.0±0.6	99.2±0.6	99.4±0.5
CA3	65.1±5.3	82.1±4.2	89.6±2.8	93.4±2.0	95.2±1.6	96.5±1.3	97.7±1.0	98.1±0.8	98.5±0.8
CA4	68.1±5.4	84.0±3.8	91.2±2.5	94.5±1.8	95.9±1.5	97.1±1.2	98.3±0.8	98.6±0.7	98.8±0.6

In Table 2~Table 5, the classification accuracy is calculated by using checking data. We prepare 100 different sets of training data for each sampling rate. We conducted 100 experiments for each proposed approach at a sampling rate. Each entry of Tables shows the average *ave* and the standard deviation *dev* in the form of *ave* ± *dev*. Mark * implies that the average of the accuracy of the proposed approach is not significantly different from that of MLEM2 by the paired t-test with significance level $\alpha = 0.05$.

From Table 2~Table 5, CA2 is significantly better than ML for all sampling rates. However, CA1, CA3 and CA4 are worse than or equal to ML, especially CA3 is significantly worse than ML. CA2 is better than CA1 and CA4 is better than CA3 at averages of accuracy. Note that CA2 and CA4 use the alleviating approach to downward unions while CA1 and CA3 apply the refining approach. Moreover, CA2 is better than CA4 and CA1 is better than CA3. CA2 and CA1 use refining approach to upward unions. Combining this observation with the data generation in the experiments, alleviating approach seems to be good for reverse-hill/reverse-valley structures, where a reverse-hill (resp. reverse-valley) structure is a structure composed of complements of upward (resp. downward) unions when upward (resp. downward) unions form a hill (resp. valley) structure. Moreover, refining approach seems to be good for hill/valley structures.

4.3 Experiments by Hill-Valley-structured Dataset

4.3.1 Dataset Generation

To obtain the hill-valley-structured dataset, we randomly generate several sequence of hyper rectangles H_l, $l = t, t+1, \cdots, p$ on condition attribute space $\times_{a \in C} V_a$ having a hill structure $H_t \supseteq H_{t+1} \supseteq \cdots \supseteq H_p$ and H_l, $l = 1, 2, \cdots, t$ having a valley structure $H_t \supseteq H_{t-1} \supseteq \cdots \supseteq H_1$ (see Figure 3).

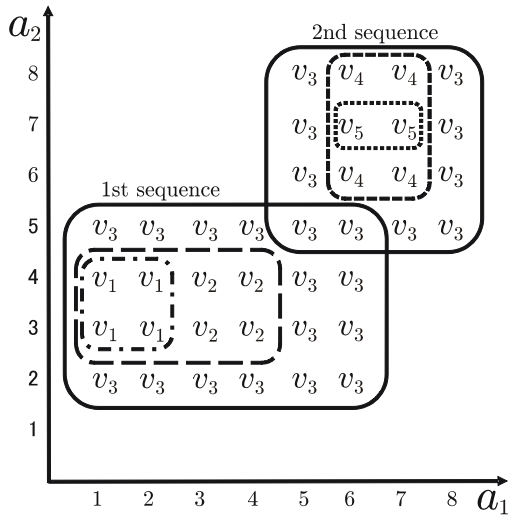

Figure 3. An example of hill-valley structure.

Then we define the decision attribute values by rules "if $x \in H_i$ and $x \notin H_{i+1}$ then decision attribute value of x is v_i" for $H_i \supseteq H_{i+1}$ and "if $x \in H_i$ and $x \notin H_{i-1}$ then decision attribute value of x is v_i" for $H_i \supseteq H_{i-1}$. Here we assume that $v_1 \leq v_2 \leq \cdots \leq v_p$, $H_0 = H_{p+1} = \varnothing$ and $\times_{a \in C} V_a \subseteq \aleph^m$ ($\aleph$ is a set of natural numbers) is finite. When x satisfies conditions of multiple rules, the decision attribute value is determined in the following procedure: (a) If x satisfies conditions of rules from hill structures, the largest value $v^{\max}$ among v_i's in their conclusions is calculated. (b) If x satisfies conditions of rules from valley structures, the smallest value $v^{\min}$ among v_i's in their conclusions is calculated. (c) If one of $v^{\max}$ and $v^{\min}$ is not calculated, the decision attribute value of x is defined by the calculated one. (d) Otherwise, we consider the latest fired rule. If the rule is defined from a hill structure, we adopt $v^{\max}$ for the decision attribute value of x, and otherwise we adopt $v^{\min}$.

4.3.2 Approach Based on the First Stage Results

In the first stage experiments, we observed that the best combination adopts the refining approach for hill/valley structures and adopts the coarsening approach for reverse-hill/reverse-valley structures of given dataset. From this observation, for dataset with more general structure, the following approach to rule induction with respect to upward unions is conceivable. (0) we select the standard value $t \in \{1, 2, \cdots, p\}$, appropriately. (1) we induce rules with respect to $Cl_t^{\geq}$ using MLEM2, (2) we induce rules with respect to $Cl_{t+1}^{\geq}$, $\cdots$, $Cl_p^{\geq}$ by the refining approach, and (3) we induce rules with respect to $Cl_{t-1}^{\geq}$, $\cdots$, $Cl_2^{\geq}$ using the coarsening approach.

To rule induction with respect to downward unions, the similar approach is conceivable under the standard value $t \in \{1, 2, \cdots, p\}$ is selected. (1) we induce rules with respect to $Cl_t^{\leq}$ using MLEM2, (2) we induce rules with respect to $Cl_{t+1}^{\leq}$, $\cdots$, $Cl_{p-1}^{\leq}$ by the coarsening approach, and (3) we induce rules with respect to $Cl_{t-1}^{\leq}$, $\cdots$, $Cl_1^{\leq}$ using the refining approach.

This approach is the mixture of refining and coarsening approaches and called the standard value switching approach. In the second stage experiments, in order to confirm the correctness of the observation, we compare the standard value switching approach with four combinations described in Table 1 and the conventional approach when the standard value t is known. The estimation of the standard value t remains for our future research.

4.3.3 Results and Discussion

We generated 40 datasets varying the numbers of condition attribute values and the number of decision classes. Both numbers of condition attributes and sequences of hyper rectangle are fixed as 4. Two of the sequences are hill-structured sequences while the other two sequences are valley-structured sequences. All condition attributes are assumed to be ordinal.

Training data are selected by random sampling from the generated datasets. The sampling size is determined by the ratio to the number of data in the generated dataset. Different ratios are considered. They are 1%, 2%, $\cdots$, 9%, 10%, 20%, $\cdots$, 70%, more concretely, the sample size is determined by [(ratio)× (the size of a dataset)], where [r] means the largest integer not greater than r.

One hundred different training data sets were prepared for each sampling size and for each data set. Then, we conducted 100 experiments using those training data sets. From each experiment, we calculated the classification accuracy using the checking data set. Moreover, we recorded the number of induced rules and the average length (the average number of condition attributes) of conditions in the premises of induced rules.

Because the observed properties of all results are similar, we show the results for two data sets, Dataset1 and Dataset 2 in Table 6~Table 9: Dataset1 has eight values for condition attributes and five values for decision attributes while Dataset 2 has eight values for condition attributes and seven values for decision attributes. In those tables, each entry $ave \pm dev$ shows the average ave and the standard deviation dev. In Table 6 and Table 7, mark * implies that the average of the accuracy of the proposed approach is not significantly different from that of MLEM2 by the paired t-test with significance level α 3 0.05.

As shown in Table 6 and Table 7, the standard value switching approach induces good sets of decision rules. From ratio 2% to 10%, the sets of decision rules induced by the standard value switching approach (SW) are significantly better than the conventional approach (ML) of MLEM2 in the classification accuracy. On the other hand, the other four approaches (CA1~CA4) are not very advantageous.

The similar results are obtained in the same experiments using other 28 data sets. Generally speaking, the standard value switching approach (SW) is signifi-

Table 6. Classification accuracy in Dataset1.

Rate	2%	3%	4%	5%	6%	8%	10%	30%
ML	57.3±5.5	66.5±5.6	73.9±4.8	78.5±4.6	82.2±4.2	88.1±4.0	91.5±2.7	98.9±0.7
CA1	55.4±6.4	64.2±5.7	72.0±7.0	76.1±5.7	79.9±4.7	85.4±4.6	89.9±3.5	98.5±0.9
CA2	55.6±6.0	64.6±5.6	72.2±6.6	75.9±5.7	80.0±5.0	85.9±4.6	89.8±3.4	98.5±0.9
CA3	55.8±6.1	65.0±6.1	73.1*±5.0	78.1*±4.9	82.0*±4.9	87.6*±4.5	91.9*±3.2	99.2±0.7
CA4	56.3*±5.4	65.1±5.9	72.9±5.4	78.0*±5.1	82.3*±4.8	88.1*±4.2	91.6*±3.1	99.1±0.7
SW	58.1*±6.6	67.6±5.9	74.9±5.5	79.2±5.1	83.2±4.4	88.9±4.4	92.7±2.6	99.2±0.7

Table 7. Classification accuracy in Dataset2.

Rate	2%	3%	4%	5%	6%	8%	10%	30%
ML	46.7±6.2	56.0±6.1	63.2±6.1	69.1±6.0	73.8±5.8	81.2±4.6	86.3±3.3	98.2±1.2
CA1	45.9*±6.6	55.5*±6.5	62.4*±6.3	67.9±6.7	72.1±6.1	79.8±5.5	84.5±4.3	97.8±1.5
CA2	46.2*±7.1	55.9*±6.5	63.0*±6.7	68.2*±6.3	72.0±6.0	79.3±5.3	83.8±3.9	97.5±1.6
CA3	45.2±5.5	54.0±6.3	61.3±7.1	68.1*±6.9	73.4*±6.5	82.0*±5.7	87.9±3.7	98.4*±1.0
CA4	45.0±6.3	54.0±5.9	61.2±7.0	67.6±6.3	72.6*±6.8	81.2*±5.2	86.3*±4.1	98.1*±1.2
SW	48.1±6.0	57.9±6.1	66.5±7.0	72.7±7.0	76.8±5.9	84.4±4.8	89.0±3.4	98.6±1.0

Table 8. The number of induced rules(left-hand side: Dataset1, right-hand side: Dataset2).

Rate	2%	4%	6%	10%	2%	4%	6%	10%
ML	24.9±2.6	32.2±2.5	35.9±2.4	39.2±2.1	34.9±4.1	47.8±4.0	53.9±5.1	59.8±3.4
CA1	28.1±3.6	37.2±4.6	42.1±4.6	44.2±3.7	40.2±5.9	57.6±7.5	65.1±8.2	72.3±8.3
CA2	29.9±4.1	38.2±4.0	42.3±4.3	44.6±4.1	43.6±6.6	60.0±7.3	66.9±7.0	76.5±7.9
CA3	29.5±3.6	37.9±4.4	40.6±4.4	42.4±2.7	43.4±6.0	58.5±7.7	64.4±7.4	66.2±5.6
CA4	31.2±3.8	38.9±4.5	40.7±4.0	42.8±3.2	46.7±7.1	60.9±7.3	66.2±6.8	70.4±6.0
SW	26.4±2.8	33.5±2.9	37.4±3.0	40.1±2.6	38.1±5.8	50.9±5.0	56.9±5.6	61.6±4.5

Table 9. The average length of conditions (left-hand side: Dataset1, right-hand side: Dataset2).

Rate	2%	4%	6%	10%	2%	4%	6%	10%
ML	2.01±0.16	2.19±0.12	2.25±0.10	2.26±0.08	1.82±0.14	2.05±0.11	2.09±0.09	2.14±0.08
CA1	2.32±0.19	2.44±0.18	2.49±0.14	2.44±0.13	2.20±0.24	2.40±0.20	2.42±0.18	2.40±0.15
CA2	2.12±0.17	2.28±0.14	2.38±0.12	2.39±0.13	1.93±0.21	2.18±0.17	2.21±0.16	2.26±0.12
CA3	2.07±0.18	2.25±0.16	2.28±0.13	2.23±0.10	1.97±0.19	2.22±0.18	2.26±0.15	2.25±0.13
CA4	1.90±0.15	2.09±0.11	2.17±0.09	2.18±0.08	1.75±0.17	2.01±0.13	2.05±0.10	2.10±0.09
SW	2.12±0.18	2.27±0.14	2.28±0.12	2.24±0.09	1.93±0.16	2.12±0.13	2.15±0.12	2.16±0.09

cantly better than the conventional approach (ML) of MLEM2 in many cases. The other four approaches (CA1~CA4) are not very advantageous. In some cases, they are better but in some other cases, they are worse than the conventional approach (ML). We cannot observe any tendency in the advantages of the four approaches (CA1~CA4).

From Table 8, we know that the number of rules induced by the standard value switching approach (SW) is smaller than the those by four approaches (CA1~CA4) but a little larger than that by the conventional approach (ML). Considering the independent applications of MLEM2, it is not very surprising that the number of rules induced by the conventional approach (ML) is the smallest.

From Table 9, we know that the average length of conditions of rules induced by the standard value switching approach (SW) is not very small. The average length of conditions of rules of the conventional approach (ML) is the smallest because of the independent applications of MLEM2.

By the above results, we have observed that, when the data are distributed along hills and valleys around a value v_t, the switching approach based on v_t is appropriate. This observation is coincident with the previous observation in the first stage experiments. Then we have confirmed that the best combination adopts the refining approach for hill/valley structures and adopts

the coarsening approach for reverse-hill/reverse-valley structures of given dataset.

5. CONCLUDING REMARKS

In this paper, we investigated the rule induction from decision tables with the ordinal condition and decision attributes. Because of the order of decision attribute values, inducing rules with respect to upward and downward unions is advantageous and then conclusions of rules may have inclusion relations. Two basic approaches for rule induction in consideration of inclusion relations of conclusions were proposed: a refining approach and a coarsening approach. Four combinations of those approaches were described. Moreover, based on the observation in the results of hill-structured datasets, the standard value switching approach was proposed under the assumptions that the standard value is known and that data are distributed along hills and valleys around the standard value. By the numerical experiments using hill-valley-structured datasets, we have observed that the standard value switching approach performed the best. Then we have concluded that the best combination adopts the refining approach for hill/valley structures and adopts the coarsening approach for reverse-hill/reverse-valley structures of given dataset.

In the real world, we may often come across data distributed along hills and valleys around some standard value. Then the standard value switching approach would be useful. However, to use this approach, we should know the standard value. The estimation of the standard value is one of the future topic of our research.

REFERENCES

Fürnkranz, J. (1999), Separate-and-conquer rule learning, *Artificial Intelligence Review*, **13**, 3-54.

Greco, S., Matarazzo, B., and Słowiński, R. (1998), New developments in the rough set approach to multi-attribute decision analysis, *Bulletin of International Rough Set Society*, **2**, 57-87.

Greco, S., Matarazzo, B., and Słowiński, R. (1999), Rough approximation of a preference relation by dominance relation, *European Journal of Operational Research*, **117**, 63-83.

Greco, S., Matarazzo, B., and Słowiński, R. (2001), Rough sets theory for multicriteria decision analysis, *European Journal of Operational Research*, **129**, 1-47.

Greco, S., Matarazzo, B., and Słowiński, R. (2002), Rough approximation by dominance relations, *International Journal of Intelligent Systems*, **17**, 153-171.

Grzymala-Busse, J. W. (1992), *LERS-A system for learning from examples based on rough sets, In R. Słowiński (ed.), Intelligent Decision Support*: *Handbook of Application and Advances of the Rough Set Theory (Dordrecht:Kluwer Academic Publishers)*, 3-18.

Grzymala-Busse, J. W. (2003), *MLEM2-discretization during rule induction*, *In M. A. Klopotek*, *S. T. Wierzchon*, *K. Trojanowski (eds.), Proceedings of the International IIS: IIPWM'03 Conference, Zakopane*, *Poland (Berlin Heidelberg*: *Springer-Verlag)*, 499-508.

Hoover, K. D. and Perez, S. J. (1999), Data mining reconsidered: encompassing and the general-to-specific approach to specification search, *Econometrics Journal*, **2**, 167-191.

Wikipedia Rule Induction (web) http://en.wikipedia.org/wiki/Rule_induction.

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

Pawlak, Z. (1991), *Rough sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishing, Dordrecht.

Stefanowski, J. (1988), Rough set based rule induction techniques for classification problems, *In Proceedings of 6th European Congress on Intelligent Techniques and Soft Computing*, *Aachen*, *Germany*, **1**, 109-113.