Extraction of Hierarchical Decision Rules from Clinical Databases using Rough Sets

Shusaku Tsumoto

Department of Medical Informatics, Shimane Medical University, Izumo, Japan
Tel: +81-853-20-2172, Fax: +81-853-20-2170, E-mail: tsumoto@computer.org

Abstract

One of the most important problems on rule induction methods is that they cannot extract rules, which plausibly represent experts' decision processes. On one hand, rule induction methods induce probabilistic rules, the description length of which is too short, compared with the experts' rules. On the other hand, construction of Bayesian networks generates too lengthy rules. In this paper, the characteristics of experts' rules are closely examined and a new approach to extract plausible rules is introduced, which consists of the following three procedures. First, the characterization of decision attributes (given classes) is extracted from databases and the classes are classified into several groups with respect to the characterization. Then, two kinds of sub-rules, characterization rules for each group and discrimination rules for each class in the group are induced. Finally, those two parts are integrated into one rule for each decision attribute. The proposed method was evaluated on a medical database, the experimental results of which show that induced rules correctly represent experts' decision processes.

Keywords:

Data Mining; Knowledge Discovery; Rule Induction; Hierarchical Decision Rules

Introduction

One of the most important problems in developing expert systems is knowledge acquisition from experts[1]. In order to automate this problem, many inductive learning methods, such as induction of decision trees[2], rule induction methods[3] and rough set theory[4,5,6,7,8,9], are introduced and applied to extract knowledge from databases, and the results show that these methods are appropriate.

However, it has been pointed out that conventional rule induction methods cannot extract rules, which plausibly represent experts' decision processes[7,8]: the description length of induced rules is too short, compared with the experts' rules. For example, rule induction methods, including AQ15[3] and PRIMEROSE[5], induce the following common rule for muscle contraction headache from databases on differential diagnosis of headache[7,8]:

[location=whole] & [Jolt Headache=no] &
[Tenderness of M1=yes] => muscle contraction headache.

This rule is shorter than the following rule given by medical experts.

[Jolt Headache=no] & [Tenderness of M1=yes]
& [Tenderness of B1=no] & [Tenderness of C1=no]
=> muscle contraction headache,

where [Tenderness of B1=no] and [Tenderness of C1=no] are added. These results suggest that conventional rule induction methods do not reflect a mechanism of knowledge acquisition of medical experts.

In this paper, the characteristics of experts' rules are closely examined and a new approach to extract plausible rules is introduced, which consists of the following three procedures. First, the characterization of each decision attribute (a given class), a list of attribute-value pairs the supporting set of which covers all the samples of the class,is extracted from databases and the classes are classified into several groups with respect to the characterization. Then, two kinds of sub-rules, rules discriminating between each group and rules classifying each class in the group are induced. Finally, those two parts are integrated into one rule for each decision attribute. The proposed method is evaluated on a medical database, the experimental results of which show that induced rules correctly represent experts' decision processes.

Background: Problems with Rule Induction

As shown in the introduction, rules acquired from medical experts are much longer than those induced from databases the decision attributes of which are given by the same experts. This is because rule induction methods generally search for shorter rules, compared with decision tree induction. In the case of decision tree induction, the induced trees are sometimes too deep and in order for the

trees to be learningful, pruning and examination by experts are required. One of the main reasons why rules are short and decision trees are sometimes long is that these patterns are generated only by one criteria, such as high accuracy or high information gain. The comparative study in this section suggests that experts should acquire rules not only by one criteria but by the usage of several measures. Those characteristics of medical experts' rules are fully examined not by comparing between those rules for the same class, but by comparing experts' rules with those for another class. For example, a classification rule for muscle contraction headache is given by:

[Jolt Headache=no] &
([Tenderness of M0=yes] or
[Tenderness of M1=yes] or
[Tenderness of M2=yes]) &
([Tenderness of B1=no] &
[Tenderness of B2=no] &
[Tenderness of B3=no] &
[Tenderness of C1=no] &
[Tenderness of C2=no] &
[Tenderness of C3=no] &
[Tenderness of C4=no])
=> muscle contraction headache.

This rule is very similar to the following classification rule for disease of cervical spine:

[Jolt Headache=no] &
([Tenderness of M0=yes] or
[Tenderness of M1=yes] or
[Tenderness of M2=yes]) &
([Tenderness of B1=yes] or
[Tenderness of B2=yes] or
[Tenderness of B3=yes] or
[Tenderness of C1=yes] or
[Tenderness of C2=yes] or
[Tenderness of C3=yes] or
[Tenderness of C4=yes])
=> disease of cervical spine.

The differences between these two rules are attribute-value pairs, from tenderness of B1 to C4. Thus, these two rules can be simplified into the following form:

$$a_1 \ \& \ A_2 \ \& \ \neg A_3 \rightarrow \textit{muscle contraction headache}$$
$$a_1 \ \& \ A_2 \ \& \ A_3 \rightarrow \textit{disease of cervical spine}$$

The first two terms and the third one represent different reasoning. The first and second term a_1 and A_2 are used to differentiate muscle contraction headache and disease of cervical spine from other diseases. The third term A_3 is used to make a differential diagnosis between these two diseases. Thus, medical experts firstly selects several diagnostic candidates, which are very similar to each other, from many diseases and then make a final diagnosis from those candidates. In the next section, a new approach for inducing the above rules is introduced.

Methods: Rule Induction Method

Rough Set Notations

In the following sections, we use the following notations introduced by Grzymala-Busse and Skowron[5], which are based on rough set theory[4]. These notations are illustrated by a small database shown in Table 1, collecting the patients who complained of headache.

Let U denote a nonempty, finite set called the universe and A denote a nonempty, finite set of attributes, i.e., $a: U \rightarrow V_a$ for $a \in A$, where V_a is called the domain of a, respectively. Then, a decision table is defined as an information system, $A=(U, A \cup \{d\})$. For example, Table 1 is an information system with $U=\{1,2,3,4,5,6\}$ and $A=\{age, location, nature, prodrome, nausea, M1\}$ and $d=class$. For $location \in A$, $V_{location}$ is defined as {occular,lateral,whole}.

The atomic formulae over $B \subseteq A \cup \{d\}$ and V are expressions of the form $[a=v]$, called descriptors over B, where $a \in B$ and $v \in V_a$. The set $F(B,V)$ of formulas over B is the least set containing all atomic formulas over B and closed with respect to disjunction, conjunction and negation. For example, $[location=occular]$ is a descriptor of B. For each $f \in F(B,V)$, f_A denote the meaning of $f \in A$, i.e., the set of all objects in U with property f, defined inductively as follows.

1. If f is of the form $[a=v]$, then, $f_A=\{s \in U \mid a(s)=v\}$
2. $(f \wedge g)_A = f_A \cap g_A$; $(f \vee g)_A = f_A \cup g_A$; $\neg f_A = U - f_A$

For example, $f=[location=whole]$ and $f_A=\{2,4,5,6\}$.

As an example of a conjuctive formula,

$$g=[location=whole] \wedge [nausea=no]$$

is a descriptor of U and g_A is equal to $g_{location,nausea}=\{2,5\}$.

Table 1- An Example of Database

No.	Age	Location	Nature	Prodrome	nausea	M1	Class
1	50-59	Ocular	Persistent	No	no	yes	m.c.h.
2	40-49	Whole	Persistent	No	no	yes	m.c.h.
3	40-49	Lateral	Throbbing	No	yes	no	Migraine
4	40-49	Whole	Throbbing	Yes	yes	no	Migraine
5	40-49	Whole	Radiating	No	no	yes	m.c.h.
6	50-59	Whole	Persistent	No	yes	yes	Psycho

DEFINITIONS: M1: tenderness of M1, m.c.h.: muscle contraction headache, migraine:classic migraine, psycho: psychogenic headache.

Accuracy and Coverage

By the use of the framework above, classification accuracy and coverage, or true positive rate is defined as follows.

Definition

Let R and D denote a formula in $F(B,V)$ and a set of

objects which belong to a decision *d*. Classification accuracy and coverage(true positive rate) for R→ d are defined as:

$$\alpha_R(D) = \frac{|R_A \cap D|}{|R_A|} (= P(D|R)),$$
$$\kappa_R(D) = \frac{|R_A \cap D|}{|D|} (= P(R|D)),$$

where $|S|$, $\alpha_R(D)$, $\kappa_R(D)$ *and* $P(S)$ denote the cardinality of a set S, a classification accuracy of R as to classification of D and coverage (a true positive rate of R to D), and probability of S, respectively.
In the above example, when R and D are set to *[nau=1]* and *[class=migraine]*, $\alpha_R(D)$=2/3=0.67 and $\kappa_R(D)$=2/2=1.0

It is notable that $\alpha_R(D)$ measures the degree of the sufficiency of a proposition, $R \rightarrow D$, and that $\kappa_R(D)$ measures the degree of its necessity. For example, if $\alpha_R(D)$ is equal to 1.0, then R →D is true. On the other hand, if $\kappa_R(D)$ is equal to 1.0, then $D \rightarrow R$ is true. Thus, if both measures are 1.0, then R ↔D.

Probabilistic Rules

By the use of accuracy and coverage, a probabilistic rule is defined as:

$$R \xrightarrow{\alpha,\kappa} d \quad s.t. \quad R = \wedge_j [a_j = v_k],$$
$$\alpha_R(D) \geq \delta_\alpha \ and \ \kappa_R(D) \geq \delta_\kappa.$$

This rule is a kind of probabilistic proposition with two statistical measures, which is an extension of Ziarko's variable precision model(VPRS)[8]. For the above example shown in Table 1, probabilistic rules for m.c.h. are given as follows:

$$[M1 = yes] \rightarrow m.c.h. \ \alpha_R(D) = 0.75, \kappa_R(D) = 1.0$$
$$[nausea = no] \rightarrow m.c.h. \ \alpha_R(D) = 1.0 \ \kappa_R(D) = 1.0$$

where δ_α and δ_κ are set to 0.75 and 0.5, respectively

It is also notable that both a positive rule and a negative rule are defined as special cases of this rule, as shown in the next subsections.

.

Characterization

In order to model medical reasoning, a statistical measure, coverage plays an important role in modeling, which is a conditional probability of a condition (R) under the decision D(P(R|D)). Let us define a characterization set of D, denoted by L(D) as a set, each element of which is an elementary attribute-value pair R with coverage being larger than a given threshold, δ_κ. That is,

$$L_{\delta\kappa}(D) = \{[a_i = v_j] \mid \kappa_{[ai=vj]}(D) \geq \delta_\kappa\}.$$

Then, three types of relations between characterization sets can be defined as follows:

$$Independent: \ L_{\delta\kappa}(D_i) \cap L_{\delta\kappa}(D_j) = \phi$$
$$Boundary: \ L_{\delta\kappa}(D_i) \cap L_{\delta\kappa}(D_j) \neq \phi$$
$$Subcategory: \ L_{\delta\kappa}(D_i) \subseteq L_{\delta\kappa}(D_j)$$

models of reasoning about complications will be All three definitions correspond to the negative region, boundary region, and positive region[4], respectively, if a set of the whole elementary attribute-value pairs will be taken as the universe of discourse.

For the above example in Table 1, let D1 and D2 be m.c.h. and Migraine and let the threshold of the coverage is larger than 0.6. Then, since

$$L_{0.6}(m.c.h.) = \{[age = 40-49], [location = whole], [nature = persistent], [prodrome = no], [nausea = no], [M1 = yes]\}$$
$$L_{0.6}(migraine) = \{[age = 40-49], [nature = throbbing], [nausea = yes], [M1 = no]\},$$

the relation between m.c.h. and migraine is boundary type when the threshold is set to 0.6. Thus, the factors that contribute to differential diagnosis between these two are: *[location=whole], [nature=persistent],[nature=throbbing], [prodrome=no], [nausea=yes], [nausea=no], [M1=yes], [M1=no]*. In these pairs, three attributes: *nausea and M1* are very important.

On the other hand, let D1 and D2 be m.c.h. and psycho and let the threshold of the coverage is larger than 0.6. Then, since

$$L_{0.6}(psycho) = \{[age = 50-59], [location = whole], [nature = persistent], [prodrome = no], [nausea = yes], [M1 = yes]\},$$

the relation between m.c.h. and psycho is also boundary. Thus, in the case of Table 1, *age, nausea and M1* are very important factors for differential diagnosis.

This relation is dependent on the value of the threshold. If the threshold is set up to 1.0, the characterization sets are:

$$L_{1.0}(m.c.h.) = \{[prodrome = no], [nausea = no], [M1 = yes]\}$$
$$L_{1.0}(migraine) = \{[age = 40-49], [nature = throbbing], [nausea = yes], [M1 = no]\},$$
$$L_{1.0}(psycho) = \{[age = 50-59], [location = whole], [nature = persistent], [prodrome = no], [nausea = yes], [M1 = yes]\}.$$

Although their contents have been changed, the relations among three diseases are still boundary and still *age, nausea and M1* are important factors. However, it is notable that the differences between characterization are much clearer.

According to the rules acquired from medical experts, medical differential diagnosis is a focusing mechanism: first, medical experts focus on some general category of diseases, such as vascular or muscular headache. After excluding the possibility of other categories, medical experts proceed into the further differential diagnosis between diseases within a general category. In this type of reasoning, subcategory type of characterization is the most important one. However, since medical knowledge has some degree of uncertainty, boundary type with high overlapped region may have to be treated like subcategory type. To check this boundary type, we use rough inclusion measure defined below.

Rough Inclusion

In order to measure the similarity between classes with respect to characterization, we introduce a rough inclusion measure μ, which is defined as follows.

$$\mu(S,T) = \frac{|S \cap T|}{|S|}$$

It is notable that if S is a subset of T, then μ(S,T)=1.0, which shows that this relation extends subset and superset relations.

This measure is introduced by Polkowski and Skowron in their study on rough mereology[10]. Whereas rough mereology firstly applies to distributed information systems, its essential idea is rough inclusion: Rough inclusion focuses on set-inclusion to characterize a hierarchical structure based on a relation between a subset and superset. Thus, application of rough inclusion to capturing the relations between classes is equivalent to constructing rough hierarchical structure between classes, which is also closely related with information granulation proposed by Zadeh[11].

Let us illustrate how this measure is applied to hierarchical rule induction by using Table 1. When the threshold for the coverage is set to 0.6,

$$\mu(L_{0.6}(m,c,h,), L_{0.6}(migraine)) = \frac{|\{[age = 40-49]\}|}{|\{[age = 40-49],[location = whole],....\}|} = \frac{1}{6}$$

$$\mu(L_{0.6}(m,c,h,), L_{0.6}(psycho)) = \frac{|\{[location = whole],[nature = persistent],[prodrome = no]\},[M1 = yes]\}|}{|\{[age = 40-49],[location = whole],....\}|} = \frac{4}{6} = \frac{2}{3}$$

$$\mu(L_{0.6}(migraine,), L_{0.6}(psycho)) = \frac{|\{[nausea = yes]\}|}{|\{[age = 40-49],[nature = throbbing],....\}|} = \frac{1}{4}$$

These values show that the characterization set of m.c.h. is closer to that of psycho than that of migraine.

When the threshold is set to 1.0,

$$\mu(L_{1.0}(m,c,h,), L_{1.0}(migraine)) = \frac{|\{\}|}{|\{[prodrome = no],[nausea = no],[M1 = yes]\}|} = 0$$

$$\mu(L_{1.0}(m,c,h,), L_{1.0}(psycho)) = \frac{|\{[prodrome = no],[M1 = yes]\}|}{|\{[prodrome = no],[nausea = no],[M1 = yes]\}|} = \frac{2}{3}$$

$$\mu(L_{1.0}(migraine), L_{1.0}(psycho)) = \frac{|\{[nausea = yes]\}|}{|\{[prodrome = no],[nausea = no],[M1 = yes]\}|} = \frac{1}{3}$$

These values also show that the characterization of m.c.h. is closer to that of psycho.

Therefore, if the threshold for rough inclusion is set to 0.6, the characterization set of m.c.h. is roughly included by that of psycho. On the other hand, the characterization set of migraine is independent of those of m.c.h. and psycho. Thus, the differential diagnosis process consists of two process: the first process should discriminate between migraine and the group of m.c.h. and psycho. Then, the second process discriminate between m.c.h and psycho. This means that the discrimination rule of m.c.h. is composed of (discrimination between migraine and the group)+ (discrimination between m.c.h. and psycho). In the case of $L_{0.6}$, since the intersection of the characerization set of m.c.h and psycho is { *[location=whole], [nature=persistent], [prodrome=no], [M1=yes]*}, and the differences in attributes between this group and migraine is *nature, M1*. So, one of the candidates of discrimination rule is *[nature=throbbing]* $\wedge$ *[M1=no]* $\rightarrow$*migraine.* The second discrimination rule is derived from the difference between the characterizaton set of m.c.h. and psycho: So, one of the candidate of the second discrimination rule is: *[age=40-49]* $\rightarrow$*m.c.h.* or *[nausea=no]*$\rightarrow$*m.c.h.* Combining these two rules, we can obtain a diagnostic rule for m.c.h as:

$\neg$*([nature=throbbing]* $\wedge$ *[M1=no])* $\wedge$*[age=40-49]* $\rightarrow$*m.c.h.*

In the case when the threshold is set to 1.0, since the

intersection of the characerization set of m.c.h and psycho is *{[prodrome=no], [M1=yes]}*, and the differences in attributes between this group and migraine is *M1*. So, one of the candidates of discrimination rule is *[M1=no] →migraine*. The second discrimination rule is derived from the difference between the characterizaton set of m.c.h. and psycho: So, one of the candidate of the second discrimination rule is: *[nausea=no]→m.c.h.* Combining these two rules, we can obtain a diagnostic rule for m.c.h as:

$$\neg([M1=no]) \wedge [nausea=no] \rightarrow m.c.h.$$

Rule Induction

Rule induction(Figure 1) consists of the following three procedures. First, the characterization of each given class, a list of attribute-value pairs the supporting set of which covers all the samples of the class, is extracted from databases and the classes are classified into several groups with respect to the characterization set. Then, two kinds of sub-rules, rules discriminating between each group and rules classifying each class in the group are induced(Figure 2). Finally, those two parts are integrated into one rule for each decision attribute(Figure 3). This method is an extension of PRIMEROSE4 reported in [9]. In the former paper, only rigid set-inclusion relations are considered for grouping; on the other hand, rough-inclusion relations are introduced in this approach. Recent empirical comparison between set-inclusion method and rough-inclusion method shows that the latter approach outperforms the former one.

Experimental Results

The above rule induction algorithm is implemented in PRIMEROSE4.5 (Probabilistic Rule Induction Method based on Rough Sets Ver 4.5), which is implemented by using SWI-prolog on Sparc Station. This system was applied to databases on differential diagnosis of headache, whose training samples consist of 1477 samples, 20 classes and 20 attributes. This system was compared with PRIMEROSE4[9], PRIMEROSE[7], C4.5[2] , AQ15[3] with respect to the following points: length of rules, similarities between induced rules and expert's rules and performance of rules. In this experiment, length was measured by the number of attribute-value pairs used in an induced rule and Jaccard's coefficient was adopted as a similarity measure[11]. Concerning the performance of rules, ten-fold cross-validation was applied to estimate classification accuracy.

Table 2 shows the experimental results, which suggest that PRIMEROSE4.5 outperforms PRIMEROSE4(set-inclusion approach) and the other four rule induction methods and induces rules very similar to medical experts' ones.

```
procedure Rule Induction (Total Process);
var
   i: integer; M,L,R: List;   DL:List; /* A list of all classes
*/
begin
   Calculate αR(D) and κR(D) for each elementary relation
   R and each class D;
   Make a list   L(D)= { R| κR(D)=1.0}   for each class D;
    while (DL≠φ)  do
      begin
         D = first(DL);   M:= DL - D;
          while (M≠φ)  do
              begin
                D(2) = first(M);
                if   (μ(L(D),L(D(2))) ≥δμ)
                then   L2(D):= L2(D)+ D(2);
                       M:= M – D(2);
                end
            Make a new decision attribute D' for L2(D);
            DL:= DL - D;
          end
      Construct a new table (T2(D)) for L2(D).
      Construct a new table(T(D'))
                for each decision attribute D';
      Induce classification rules R2 for each L2(D)
(Fig.2 )
      Store Rules into a List R(D);
      Induce classification rules Rd for each D' in T(D');
                        (Fig.2)
      Store Rules into a List R(D')(=R(L2(Di)))
      Integrate R2 and Rd into a rule RD; /* Fig.3 */
end {Rule Induction};
```

Figure 1- An Algorithm for Rule Induction

```
procedure Induction of Classification Rules;
   var   i: integer;   M, Li: List;
   begin
    L1:= (a list of all attribute-value pairs);
    i:=1; M:={};
        for i:=1 to n do
        /* n: Total number of attributes in a database */
          begin
            while (Li ≠{}) do
              begin
                 Select one pair R=∧ [ai=vj] from Li ;
                 Li := Li - {R} ;
                 if (κR(D) ≥ δα) then do
                   if αR(D) ≥δα) then do
                      Sir := Sir ∪{R};
                    else M := M ∪{R};
              end
            if (M={}) quit.
             else
              Li+1:= (A list of the whole combination of the
                              conjunction formulae in M);
            end
    end {Induction of Probabilistic Rules}
```

Figure 2 -Induction of Classification Rules

```
procedure Rule Integration;
var
    i: integer; M, L2: List;
    R(D): List; /* A list of rules for D */
    DL: List; /* A list of all classes */
begin
    while (DL≠ϕ)  do
          begin
              D = first(DL);   M:= DL - D;
              Select one rule $R'→ D' from R(L2(D));
              while (M≠ϕ)  do
                   begin
                     D2:= first(M);\
                     Select one rule $R→ D2 for D2;
                     Integrate two rules: R & R' → D2;
                     M:= M-D2;
                  end
                   DL:= DL-D;
              end
  end {Rule Integration}
```

Figure 3- An Algorithm for Rule Integration

Table 2- Experimental Results

Method	Length	Similarity	Accuracy
PRIMEROSE4.5	8.8±0.27	0.95±0.08	0.952±0.027
PRIMEROSE4.0	8.6±0.27	0.93±0.08	0.933±0.027
C4.5	3.9±0.39	0.37±0.12	0.758±0.019
AQ15	4.7±0.36	0.49±0.12	0.793±0.03
Experts	9.1±0.33	1.00±0.00	0.972±0.014

Discussion

Focusing Mechanism

One of the most interesting features in medical reasoning is that medical experts make a differential diagnosis based on focusing mechanisms: with several inputs, they eluminate some candidates and proceed into further steps. In this elimination, our empirical results suggest that grouping of diseases are very important to realize automated acquisition of medical knowledge from clinical databases. Readers may say that conceptual clustering or nearest neighborhood methods[12] will be useful for grouping. However, those two methods are based on classification accuracy, that is, they induce grouping of diseases, whose rules are of high accuracy. Their weak point is that they do not reflect medical reasoning: focusing mechanisms of medical experts are chiefly based not on classification accuracy, but on coverage. Thus, we focus on the role of coverage in focusing mechanisms and propose an algorithm on grouping of diseases by using this measure. The above experiments show that rule induction with this grouping generates rules, which are similar to medical experts' rules and they suggest that our proposed method should capture medical experts' reasoning.

Exclusive Rules

An exclusive rule is defined as a rule supported by all the positive examples, the coverage of which is equal to 1.0. It is notable that an exclusive rule represents the necessity condition of a decision and also that the set supporting a exclusive rule corresponds to the upper approximation of a target concept, which is introduced in rough sets[3]. Thus, an exclusive rule is represented as:

$$R \xrightarrow{\alpha,\kappa} d \quad s.t.\ R = \vee_j [a_j = v_k],\ \kappa_R(D) = 1.0,$$

In the above example, exclusive rule of ``m.c.h." is:
[M1=yes] ∨[nau=no] →m.c.h. κ=1.0,
It is notable that from the viewpoint of propositional logic, an exclusive rule should be represented as:

$$d \rightarrow \vee_j [a_j = v_k] \quad s.t.\ \kappa_R(D) = 1.0.$$

Because the condition of an exclusive rule corresponds to the necessity condition of conclusion *d*. Thus, it is easy to see that a negative rule is defined as the contrapositive of an exclusive rule:

$$\wedge_j \neg[a_j = v_k] \rightarrow \neg d,$$

which means that if a case does not satisfy any attribute value pairs in the condition of a negative rules, then we can exclude a decision *d* from candidates.

Total Covering Rules (Disease Image)

Originally, a total covering rule is defined as a set of symptoms which can be observed at least in one case of a target disease. That is, this rule is defined as a collection of attribute-value pairs whose accuracy is larger than 0:

$$R \xrightarrow{\alpha,\kappa} d \quad s.t.\ R = \vee_j [a_j = v_k],\ \alpha_R(D) > 0.$$

From the definition of accuracy and coverage, this formula can be transformed into:

$$R \xrightarrow{\alpha,\kappa} d \quad s.t.\ R = \vee_j [a_j = v_k],\ \kappa_R(D) > 0.$$

For each attribute, the attribute-value pairs form a partition of *U*. Thus, for each attribute, total covering rules include a covering of all the positive examples. According to this property, the above formula is redefined as:

$$R \xrightarrow{\alpha,\kappa} d \quad s.t.\ R = \vee_j R(a_j),$$
$$R(a_j) = \vee_k [a_j = v_k]\ s.t.\ \kappa_R(D) = 1.0.$$

It is notable that this definition is an extension of exclusive rules and this total covering rule can be written as:

$$d \rightarrow \vee_j \vee_k [a_j = v_k] \quad s.t.\ \kappa_{[a_j = v_k]}(D) > 0.$$

Relations between Rules

Let *S(R)* denote a set of attribute-value pairs of rule *R*.

For each class d, let R_{pos} (d), R_{ex} (d) and R_{tc}(d) denote the positive, exclusive rule and total covering rules, respectively. Then, since

$$S(R_{ex}(d)) \subseteq S(R_{tc}(d)),$$

a total covering rule can be viewed as an upper approximation of exclusive rules. It is also notable that the above relation will hold in the relation between positive rule(inclusive rule) and total covering rule. That is,

$$S(R_{pos}(d)) \subseteq S(R_{tc}(d)),$$

Thus, total covering rule can be viewed as an upper approximation of inclusive rules. This relation also holds when R_{pos} (d) is replaced with a probabilistic rule, which shows that total covering rules is the weakest form of diagnostic rules.

Extension of Exclusive Rules

The definition of exclusive rule is very strict because each attribute-value pair covers all the positive examples. This strict condition can be relaxed if the disjunction of attribute-value pairs is allowed. That is,

$$R \xrightarrow{\alpha,\kappa} d \quad s.t. \quad R = \vee_j R(a_j),$$
$$R(a_j) = \vee_{i<k(j)} [a_j = v_i] \; s.t. \; \kappa_R(D) = 1.0,$$

where $k(j)$ denotes the total number of values in attribute a_j

Let R_{ex2} (d) denote this type of extension of the exclusive rule.

Then, the following relation is obtained:

$$S(R_{ex}(d)) \subseteq S(R_{ex2}(d)) \subseteq S(R_{tc}(d)),$$

Thus, the strict exclusive rule and total covering rule can be viewed as an lower and upper approximation of the extended exclusive rules.

These inclusion-exclusion relations suggest that such a set-theoretical relations are implicitly implemented in the reasoning of medical reasoning. It will be our future work to investigate these characteristics further, which will show the cognitive aspects of rough set theory.

Conclusion

In this paper, the characteristics of experts' rules are closely examined, whose empirical results suggest that grouping of diseases are very important to realize automated acquisition of medical knowledge from clinical databases. Thus, we focus on the role of coverage in focusing mechanisms and propose an algorithm on grouping of diseases by using this measure. The above experiments show that rule induction with this grouping generates rules, which are similar to medical experts' rules and they suggest that our proposed method should capture medical experts' reasoning. The proposed method was evaluated on three medical databases, the experimental results of which show that induced rules correctly represent experts' decision processes.

Acknowledgments

The author thank Professor T.Y. Lin and Andrzej Skowron for the insightful discussions.

This work was supported by the Grant-in-Aid for Scientific Research on Priority Areas(B) (No.759) "Implementation of Active Mining in the Era of Information Flood" by the Ministry of Education, Science, Culture, Science and Technology of Japan.

References

[1] Buchnan BG and Shortliffe EH. *Rule-Based Expert Systems*, Addison-Wesley, New York, 1984.

[2] Quinlan JR. *C4.5 - Programs for Machine Learning*, Morgan Kaufmann, Palo Alto, 1993.

[3] Michalski RS, Mozetic I, Hong J, and Lavrac N. The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains, in: *Proceedings of the fifth National Conference on Artificial Intelligence*, AAAI Press, Menlo Park, 1986: pp. 1041-1045.

[4] Pawlak Z. *Rough Sets*. Kluwer Academic Publishers, Dordrecht, 1991.

[5] A. Skowron, A. and J. Grzymala-Busse, "From rough set theory to evidence theory". In: Yager, R., Fedrizzi, M. and Kacprzyk, J.(eds.) Advances in the Dempster-Shafer Theory of Evidence, John Wiley & Sons, NY, 1994: 193-236.

[6] Tsumoto S and Tanaka H. PRIMEROSE: Probabilistic Rule Induction Method based on Rough Sets and Resampling Methods. *Computational Intelligence* 1995:11, 389-405.

[7] Ziarko W. Variable Precision Rough Set Model. *Journal of Computer and System Sciences*1993:46, 39-59.

[8] Tsumoto S. Automated Induction of Medical Expert System Rules from Clinical Databases based on Rough Set Theory. *Information Sciences* 1998:112, 67-84.

[9] Tsumoto S. Extraction of Experts' Decision Rules from Clinical Databases using Rough Set Model. *Journal of Intelligent Data Analysis* 1998: 2(3).

[10]Polkowski L. and Skowron A.: Rough mereology: a new paradigm for approximate reasoning. *Intern. J. Approx. Reasoning* 1996: 15, 333--365.

[11]Zadeh LA. Toward a theory of fuzzy information granulation and its certainty in human reasoning and fuzzy logic. *Fuzzy Sets and Systems* 1997: 90, 111-127.

[12] Everitt BS. *Cluster Analysis*, 3rd Edition, John Wiley & Son, London, 1996.