Information Refinement and Revision for Medical Expert System – Automated Extraction of Hierarchical Rules from Clinical Data –

Shusaku Tsumoto
Department of Medicine Informatics,
Shimane Medical University, School of Medicine,
89-1 Enya-cho Izumo City, Shimane 693-8501 Japan
E-mail: tsumoto@computer.org

Abstract

Since real-world decision making include several decision steps, real-world decision making agents have several sophisticated diagnostic reasoning mechanisms. Thus, if we want to update these reasoning steps, we have to extract rules for each step from real-world datasets. However, one of the most important problems on rule induction methods is that they aim at induction of simple rules and cannot extract rules that plausibly represent experts decision processes, which makes rule induction methods not applicable to the maintenance of real-world decision making agents. 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 medical databases, the experimental results of which show that induced rules correctly represent experts decision processes.

Introduction

One of the most important problems in developing expert systems is knowledge acquisition from experts (Buchanan and Shortliffe, 1984). In order to automate this problem, many inductive learning methods, such as induction of decision trees (Breiman, 1984; Quinlan, 1993), rule induction methods (Michalski, 1986; Shavlik, 1990) and rough set theory (Pawlak, 1991; Tsumoto, 1995; Ziarko, 1993), 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(Tsumoto 1995; Tsumoto 1998): the description length of induced rules is too short, compared with the experts' rules. For example, rule induction methods, including AQ15(Michalski 1986) and PRIMEROSE(Tsumoto 1995) induce the following common rule for muscle contraction headache from databases on differential diagnosis of headache(Tsumoto 1998):

$$\begin{aligned} [location = whole] \quad & \wedge [\text{Jolt Headache} = no] \\ & \wedge [\text{Tenderness of M1} = yes] \\ & \rightarrow \text{muscle contraction headache.} \end{aligned}$$

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

$$\begin{aligned} & [\text{Jolt Headache} = no] \\ & \wedge([\text{Tenderness of M0} = yes] \\ & \quad \vee[\text{Tenderness of M1} = yes] \\ & \quad \vee[\text{Tenderness of M2} = yes]) \\ & \wedge[\text{Tenderness of B1} = no] \\ & \wedge[\text{Tenderness of B2} = no] \\ & \wedge[\text{Tenderness of B3} = no] \\ & \wedge[\text{Tenderness of C1} = no] \\ & \wedge[\text{Tenderness of C2} = no] \\ & \wedge[\text{Tenderness of C3} = no] \\ & \wedge[\text{Tenderness of C4} = no] \\ & \rightarrow \text{muscle contraction headache} \end{aligned}$$

where $[\text{Tenderness of B1} = no]$ and $[\text{Tenderness of C1} = no]$ 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 medical databases, 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:

$$
\begin{array}{l}
[\text{Jolt Headache} = no] \\
\wedge([\text{Tenderness of M0} = yes] \\
\quad \vee[\text{Tenderness of M1} = yes] \\
\quad \vee[\text{Tenderness of M2} = yes]) \\
\wedge[\text{Tenderness of B1} = no] \\
\wedge[\text{Tenderness of B2} = no] \\
\wedge[\text{Tenderness of B3} = no] \\
\wedge[\text{Tenderness of C1} = no] \\
\wedge[\text{Tenderness of C2} = no] \\
\wedge[\text{Tenderness of C3} = no] \\
\wedge[\text{Tenderness of C4} = no] \\
\rightarrow \text{muscle contraction headache}
\end{array}
$$

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

$$
\begin{array}{l}
[\text{Jolt Headache} = no] \\
\wedge([\text{Tenderness of M0} = yes] \\
\quad \vee[\text{Tenderness of M1} = yes] \\
\quad \vee[\text{Tenderness of M2} = yes]) \\
\wedge([\text{Tenderness of B1} = yes] \\
\quad \vee[\text{Tenderness of B2} = yes] \\
\quad \vee[\text{Tenderness of B3} = yes] \\
\quad \vee[\text{Tenderness of C1} = yes] \\
\quad \vee[\text{Tenderness of C2} = yes] \\
\quad \vee[\text{Tenderness of C3} = yes] \\
\quad \vee[\text{Tenderness of C4} = yes]) \\
\rightarrow \text{disease of cervical spine}
\end{array}
$$

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:

$$
\begin{array}{rcl}
a_1 \wedge A_2 \wedge \neg A_3 & \rightarrow & \textit{muscle contraction headache} \\
a_1 \wedge A_2 \wedge A_3 & \rightarrow & \textit{disease of cervical spine}
\end{array}
$$

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.

Table 1: An Example of Database

	age	loc	nat	prod	nau	M1	class
1	50...59	occ	per	no	no	yes	m.c.h.
2	40...49	who	per	no	no	yes	m.c.h.
3	40...49	lat	thr	yes	yes	no	migra
4	40...49	who	thr	yes	yes	no	migra
5	40...49	who	rad	no	no	yes	m.c.h.
6	50...59	who	per	no	yes	yes	psycho

DEFINITIONS: loc: location, nat: nature, prod: prodrome, nau: nausea, M1: tenderness of M1, who: whole, occ: occular, lat: lateral, per: persistent, thr: throbbing, rad: radiating, m.c.h.: muscle contraction headache, migra: migraine, psycho: psychological pain,

Rough Set Theory and Probabilistic Rules

Rough Set Notations

In the following sections, we use the following notations introduced by Skowron and Grzymala-Busse(1994), which are based on rough set theory(Pawlak, 1991). 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, $f_A = \{s \in U | a(s) = v\}$
2. $(f \wedge g)_A = f_A \cap g_A$; $(f \vee g)_A = f_A \vee g_A$; $(\neg f)_A = U - f_a$

For example, $f = [location = whole]$ and $f_A = \{2,4,5,6\}$. As an example of a conjunctive formula, $g = [location = whole] \wedge [nausea = no]$ is a descriptor of U and f_A is equal to $g_{location,nausea} = \{2,5\}$.

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

Definition 1

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 $R \rightarrow d$ is defined as:*

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

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 $R \rightarrow D$ 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 $R \leftrightarrow D$.

Probabilistic Rules

According to the definitions, probabilistic rules with high accuracy and coverage are defined as:

$$R \stackrel{\alpha,\kappa}{\rightarrow} d \quad \text{s.t.} \quad R = \vee_i R_i = \vee \wedge_j [a_j = v_k], \; \alpha_{R_i}(D) \geq \delta_\alpha \text{ and } \kappa_{R_i}(D) \geq \delta_\kappa,$$

where δ_α and δ_κ denote given thresholds for accuracy and coverage, respectively. For the above example shown in Table 1, probabilistic rules for m.c.h. are given as follows:

$$[M1 = yes] \rightarrow m.c.h. \quad \alpha = 3/4 = 0.75, \kappa = 1.0,$$
$$[nau = no] \rightarrow m.c.h. \quad \alpha = 3/3 = 1.0, \kappa = 1.0,$$

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

Characterization Sets

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,

$$L_{\delta_\kappa} = \{[a_i = v_j] | \kappa_{[a_i=v_j]}(D) \geq \delta_\kappa$$

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

Independent type: $L_{\delta_\kappa}(D_i) \cap L_{\delta_\kappa(D_j)} = \phi$,
Boundary type: $L_{\delta_\kappa}(D_i) \cap L(D_j)_{\delta_\kappa} \neq \phi$, and
Positive type: $L_{\delta_\kappa}(D_i) \subseteq L_{\delta_\kappa}(D_j)$.

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 D_1 and D_2 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 D_1 and D_2 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.

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

It is notable that if $S \subseteq T$, then $\mu(S,T) = 1.0$, which shows that this relation extends subset and superset relations. This measure is introduced by Polkowski and Skowron (1996) in their study on rough mereology. 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(1997). 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,

$$\mu(L_{0.6}(m.c.h.), L_{0.6}(migraine)) = \tfrac{1}{6}$$
$$\mu(L_{0.6}(m.c.h.), L_{0.6}(psycho)) = \tfrac{4}{6} = \tfrac{2}{3}$$
$$\mu(L_{0.6}(migraine), L_{0.6}(psycho)) = \tfrac{1}{4}$$

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,

$$\mu(L_{1.0}(m.c.h.), L_{1.0}(migraine)) = 0$$
$$\mu(L_{1.0}(m.c.h.), L_{1.0}(psycho)) = \tfrac{2}{3}$$
$$\mu(L_{1.0}(migraine), L_{1.0}(psycho)) = \tfrac{1}{4}$$

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 L0.6, 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 $[M1 = no] \rightarrow migraine$. 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: $[nausea = no] \rightarrow m.c.h.$ 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(Fig 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. Then, two kinds of sub-rules, rules discriminating between each group and rules classifying each class in the group are induced(Fig 2). Finally, those two parts are integrated into one rule for each decision attribute(Fig 3).[1]

Example

Let us illustrate how the introduced algorithm works by using a small database in Table 1. For simplicity, two thresholds δ_α and δ_μ are set to 1.0, which means that only deterministic rules should be induced and that only subset and superset relations should be considered for grouping classes.

After the first and second step, the following three sets will be obtained: $L(m.c.h.) = \{[prod = no], [M1 = yes]\}$, $L(migra) = \{[age = 40...49], [nat = who], [prod = yes], [nau = yes], [M1 = no]\}$, and $L(psycho) = \{[age = 50...59], [loc = who], [nat = per], [prod = no], [nau = no], [M1 = yes]\}$. Thus, since a relation $L(psycho) \subset L(m.c.h.)$ holds (i.e.,$\mu(L(m.c.h.), L(psycho)) = 1.0$), a new decision attribute is $D_1 = \{m.c.h., psycho\}$ and $D_2 = \{migra\}$, and a partition $P = \{D_1, D_2\}$ is obtained. From this partition, two decision tables will be generated, as shown in Table 2 and Table 3 in the fifth step.

In the sixth step, classification rules for D_1 and D_2 are induced from Table 2. For example, the following rules are obtained for D_1.

[1]This method is an extension of PRIMEROSE4 reported in (Tsumoto, 1998b). 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.

procedure $Rule\ Induction\ (Total\ Process)$;
var
$i : integer; \quad M, L, R : List;$
$L_D : List$; /* A list of all classes */
begin
Calculate $\alpha_R(D_i)$ and $\kappa_R(D_i)$
for each elementary relation R and each class D_i;
Make a list $L(D_i) = \{R|\kappa_R(D) = 1.0\})$
for each class D_i;
while ($L_D \neq \phi$) **do**
begin
$D_i := first(L_D)$; $M := L_D - D_i$;
while ($M \neq \phi$) **do**
begin
$D_j := first(M)$;
if ($\mu(L(D_j), L(D_i)) \leq \delta_\mu$)
then $L_2(D_i) := L_2(D_i) + \{D_j\}$;
$M := M - D_j$;
end
Make a new decision attribute D'_i for $L_2(D_i)$;
$L_D := L_D - D_i$;
end
Construct a new table ($T_2(D_i)$)for $L_2(D_i)$.
Construct a new table($T(D'_i)$)
for each decision attribute D'_i;
Induce classification rules R_2 for each $L_2(D)$;
/* Fig.2 */
Store Rules into a List $R(D)$
Induce classification rules R_d
for each D' in $T(D')$; /* Fig.2 */
Store Rules into a List $R(D')(= R(L_2(D_i)))$
Integrate R_2 and R_d into a rule R_D; /* Fig.3 */
end $\{Rule\ Induction\ \}$;

Figure 1: An Algorithm for Rule Induction

$[M1 = yes] \quad \to D_1 \quad \alpha = 1.0, \kappa = 1.0,$ supported by {1,2,5,6}
$[prod = no] \quad \to D_1 \quad \alpha = 1.0, \kappa = 1.0,$ supported by {1,2,5,6}
$[nau = no] \quad \to D_1 \quad \alpha = 1.0, \kappa = 0.75,$ supported by {1,2,5}
$[nat = per] \quad \to D_1 \quad \alpha = 1.0, \kappa = 0.75,$ supported by {1,2,6}
$[loc = who] \quad \to D_1 \quad \alpha = 1.0, \kappa = 0.75,$ supported by {2,5,6}
$[age = 50...59] \quad \to D_1 \quad \alpha = 1.0, \kappa = 0.5,$ supported by {2,6}

In the seventh step, classification rules for $m.c.h.$ and $psycho$ are induced from Table 3. For example, the following rules are obtained from $m.c.h.$.

$[nau = no] \quad \to m.c.h. \quad \alpha = 1.0, \kappa = 1.0,$ supported by {1,2,5}
$[age = 40...49] \quad \to m.c.h. \quad \alpha = 1.0, \kappa = 0.67,$ supported by {2,5}

In the eighth step, these two kinds of rules are integrated in the following way. Rule $[M1 = yes] \to D_1$, $[nau = no] \to m.c.h.$ and $[age = 40...49] \to m.c.h.$ have a supporting set which is a subset of {1,2,5,6}. Thus, the following rules are obtained:

$[M1 = yes]$ & [nau=no] $\to m.c.h.$
$\alpha = 1.0, \kappa = 1.0$, supported by {1,2,5}
$[M1 = yes]$ & [age=40...49] $\to m.c.h.$
$\alpha = 1.0, \kappa = 0.67$, supported by {2,5}

procedure $Induction\ of\ Classification\ Rules$;
var
$i : integer; \quad M, L_i : List;$
begin
$L_1 := L_{er}$; /* L_{er}: List of Elementary Relations */
$i := 1; \quad M := \{\}$;
for $i := 1$ **to** n **do** /* n: Total number of attributes */
begin
while ($L_i \neq \{\}$) **do**
begin
Select one pair $R = \wedge[a_i = v_j]$ from L_i;
$L_i := L_i - \{R\}$;
if $(\alpha_R(D) \geq \delta_\alpha)$ and $(\kappa_R(D) \geq \delta_\kappa)$
then do $S_{ir} := S_{ir} + \{R\}$;
/* Include R as Inclusive Rule */
else $M := M + \{R\}$;
end
$L_{i+1} :=$ (A list of the whole combination of the conjunction formulae in M);
end
end $\{Induction\ of\ Classification\ Rules\ \}$;

Figure 2: An Algorithm for Classification Rules

Table 2: A Table for a New Partition P

	age	loc	nat	prod	nau	M1	class
1	50...59	occ	per	0	0	1	D_1
2	40...49	who	per	0	0	1	D_1
3	40...49	lat	thr	1	1	0	D_2
4	40...49	who	thr	1	1	0	D_2
5	40...49	who	rad	0	0	1	D_1
6	50...59	who	per	0	1	1	D_1

Experimental Results

The above rule induction algorithm is implemented in PRIMEROSE4.5 (Probabilistic Rule Induction Method

Table 3: A Table for D_1

	age	loc	nat	prod	nau	M1	class
1	50...59	occ	per	0	0	1	m.c.h.
2	40...49	who	per	0	0	1	m.c.h.
5	40...49	who	rad	0	0	1	m.c.h.
6	50...59	who	per	0	1	1	psycho

```
procedure Rule Integration;
  var
    i : integer;   M, L_2 : List;
    R(D_i) : List; /* A list of rules for D_i */
    L_D : List; /* A list of all classes */
  begin
    while(L_D ≠ φ) do
      begin
        D_i := first(L_D); M := L_2(D_i);
        Select one rule R' → D'_i from R(L_2(D_i)).
        while (M ≠ φ) do
          begin
            D_j := first(M);
              Select one rule R → d_j for D_j;
              Integrate two rules: R ∧ R' → d_j.
            M := M − {D_j};
          end
        L_D := L_D − D_i;
      end
  end {Rule Combination}
```

Figure 3: An Algorithm for Rule Integration

based on Rough Sets Ver 4.5), [2] and 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, PRIMEROSE, C4.5, CN2 (Clark,1989), AQ15 and k-NN (Aha, 1991) [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 (Everitt, 1996). Concerning the performance of rules, ten-fold cross-validation was applied to estimate classification accuracy.

Table 4 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.

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(k-NN) 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.

Precision for Probabilistic Rules

In the above experiments, the thresholds δ_α and δ_κ for selection of inclusive rules were set to 0.75 and 0.5, respectively. Although this precision contributes to the reduction of computational complexity, this methodology, which gives a threshold in a static way, causes a serious problem. For example, there exists a case when the accuracy for the first, the second, and the third candidate is 0.5, 0.49, and 0.01, whereas accuracy for other classes is almost equal to 0. Formally, provided an attribute-value pair, R, the following equations hold: $\alpha_R(D_1) = 0.5, \alpha_R(D_2) = 0.49, \alpha_R(D_3) = 0.01$, and $\alpha_R(D_i) \approx 0 (i = 4, \cdots, 10)$. Then, both of the first and the second candidates should be suspected because those accuracies are very close, compared with the accuracy for the third and other classes. However, if a threshold is statically set to 0.5, then this pair is not included in positive rules for D_2. In this way, a threshold should be determined dynamically for each attribute-value pair. In the above example, an attribute-value pair should be included in positive rules of D_1 and D_2.

From discussion with domain experts, it is found that this type of reasoning is very natural, which may contribute to the differences between induced rules and the ones acquired from medical experts. Thus, even in a learning algorithm, comparison between the whole given classes should be included in order to realize more plausible reasoning strategy.

Unfortunately, since the proposed algorithm runs for each disease independently, the above type of reasoning cannot be incorporated in a natural manner, which causes computational complexity to be higher. It is our future work to develop such interacting process in the learning algorithm.

Generality of the Proposed Method

One major character of medical reasoning is that medical experts finally select one or two diagnostic candidates from many diseases, called focusing mechanism. For example, in differential diagnosis of headache, experts choose one from about 60 diseases. The proposed method models induction of rules which incorporates this mechanism, whose experimental evaluation show that induced rules correctly represent medical experts' rules.

This focusing mechanism is not only specific to medical domain. In a domain in which a few diagnostic conclusions should be selected from many candiates, this mechanism can be applied. For example, fault diagnosis of complicated electronic devices should focus on which components will

[2] The program is implemented by using SWI-prolog(1995) on Sparc Station 20.

[3] The most optimal k for each domain is attached to Table 4.

Table 4: Experimental Results

Method	Length	Similarity	Accuracy
		Headache	
PRIMEROSE4.5	8.8 ± 0.27	0.95 ± 0.08	$95.2 \pm 2.7\%$
PRIMEROSE4.0	8.6 ± 0.27	0.93 ± 0.08	$93.3 \pm 2.7\%$
Experts	9.1 ± 0.33	1.00 ± 0.00	$98.0 \pm 1.9\%$
PRIMEROSE	5.3 ± 0.35	0.54 ± 0.05	$88.3 \pm 3.6\%$
C4.5	4.9 ± 0.39	0.53 ± 0.10	$85.8 \pm 1.9\%$
CN2	4.8 ± 0.34	0.51 ± 0.08	$87.0 \pm 3.1\%$
AQ15	4.7 ± 0.35	0.51 ± 0.09	$86.2 \pm 2.9\%$
k-NN (7)	6.7 ± 0.25	0.61 ± 0.09	$88.2 \pm 1.5\%$

k-NN (i) shows the value of i which gives the highest performance in k ($1 \leq k \leq 20$).

cause a functional problem: the more complicated devices are, the more sophisticated focusing mechanism is required. In such domain, proposed rule induction method will be useful to induce correct rules from datasets.

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

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

Agrawal, R., Imielinski, T., and Swami, A., Mining association rules between sets of items in large databases, in *Proceedings of the 1993 International Conference on Management of Data (SIGMOD 93)*, pp. 207-216, 1993.

Aha, D. W., Kibler, D., and Albert, M. K., Instance-based learning algorithm. *Machine Learning*, 6, 37-66, 1991.

Breiman, L., Freidman, J., Olshen, R., and Stone, C., *Classification And Regression Trees*, Wadsworth International Group, Belmont, 1984.

Buchnan, B. G. and Shortliffe, E. H., *Rule-Based Expert Systems*, Addison-Wesley, New York, 1984.

Clark, P. and Niblett, T., The CN2 Induction Algorithm. *Machine Learning*, 3, 261-283, 1989.

Everitt, B. S., *Cluster Analysis*, 3rd Edition, John Wiley & Son, London, 1996.

Michalski, R. S., 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*, 1041-1045, AAAI Press, Menlo Park, 1986.

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

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

Quinlan, J.R., *C4.5 - Programs for Machine Learning*, Morgan Kaufmann, Palo Alto, 1993.

Shavlik, J. W. and Dietterich, T.G., (eds.) *Readings in Machine Learning*, Morgan Kaufmann, Palo Alto, 1990.

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

SWI-Prolog Version 2.0.9 Manual, University of Amsterdam, 1995.

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

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

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

Zadeh, L.A., Toward a theory of fuzzy information granulation and its certainty in human reasoning and fuzzy logic. *Fuzzy Sets and Systems* **90**, 111-127, 1997.

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