Mining Rules of Multi-level Diagnostic Procedure from Databases

Shusaku Tsumoto

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

Abstract. One of the most important features of expert reasoning is that each reasoning rule may be composed of several diagnostic steps, usually hierarchical differential diagnosis. For example, medical diagnosis include hierarchical diagnostic steps In this paper, the characteristics of experts' rules are closely examined from the viewpoint of hiearchical decision steps 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 concept hierarchy for given classes is calculated. Second, based on the hierarchy, rules for each hierarchical level are induced from data. Then, for each given class, rules for all the hierarchical levels are integrated into one rule. The proposed method was evaluated on medical databases, the experimental results of which show that induced rules correctly represent experts' decision processes.

1 Introduction

One of the most important problems in data mining is that extracted rules are not easy for domain experts to interpret. One of its reasons is that conventional rule induction methods[7] cannot extract rules, which plausibly represent experts' decision processes[9]: the description length of induced rules is too short, compared with the experts' rules. For example, rule induction methods, including AQ15[4] and PRIMEROSE[9], induce the following common rule for muscle contraction headache from databases on differential diagnosis of headache:

$[location = whole] \wedge$[Jolt Headache $= no$] $\wedge$[Tenderness of M1 $= yes$]
$\rightarrow$ muscle contraction headache.

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

[Jolt Headache $= no$]
$\wedge$([Tenderness of M0 $= yes$] $\vee$[Tenderness of M1 $= yes$] $\vee$[Tenderness of M2 $= yes$])
$\wedge$[Tenderness of B1 $= no$] $\wedge$[Tenderness of B2 $= no$] $\wedge$[Tenderness of B3 $= no$]
$\wedge$[Tenderness of C1 $= no$] $\wedge$[Tenderness of C2 $= no$] $\wedge$[Tenderness of C3 $= no$]
[Tenderness of C4 $= no$] $\rightarrow$ muscle contraction headache

where [Tenderness of B1 $= no$] and [Tenderness of C1 $= no$] are added.

N. Lavrač et al. (Eds.): PKDD 2003, LNAI 2838, pp. 459–470, 2003.

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 paper is organized as follows. Section 2 discusses the background of this study. Section 3 and 4 introduces rough sets and a characterization set. Section 5 gives an algorithm for rule induction. Section 6 shows an illustrative example and Section 7 discusses the results. Finally, Section 8 concludes this paper.

2 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. One of the main reasons why rules are short is that these patterns are generated only by one criteria, such as high accuracy or high information gain. The comparative studies[9,10] suggest 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[9]. For example, the classification rule for muscle contraction headache given in Section 1 is very similar to the following classification rule for disease of cervical spine:

[Jolt Headache = *no*]
∧([Tenderness of M0 = *yes*] ∨[Tenderness of M1 = *yes*] ∨[Tenderness of M2 = *yes*])
∧([Tenderness of B1 = *yes*] ∨[Tenderness of B2 = *yes*] ∨[Tenderness of B3 = *yes*]
∨[Tenderness of C1 = *yes*] ∨[Tenderness of C2 = *yes*] ∨[Tenderness of C3 = *yes*]
∨[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:

$$A_1 \wedge A_2 \wedge \neg A_3 \rightarrow \textit{muscle contraction headache}$$
$$A_1 \wedge A_2 \wedge A_3 \rightarrow \textit{disease of cervical spine},$$

where A_1, A_2 and A_3 are given as the following formulae:
A_1 = [Jolt Headache = *no*], A_2 = [Tenderness of M0 = *yes*] ∨ [Tenderness of $M1$ = *yes*] ∨ [Tenderness of M2 = *yes*], and A_3 = [Tenderness of C1 = *no*] ∧ [Tenderness of C2 = *no*] ∧ [Tenderness of C3 = *no*] ∧ [Tenderness of C4 = *no*].

The first two blocks (A_1 and A_2) and the third one (A_3) represent the different types of differential diagnosis. The first one A_1 shows the discrimination between muscular type and vascular type of headache. Then, the second part shows that between headache caused by neck and head muscles. Finally, the third formula A_3 is used to make a differential diagnosis between muscle contraction headache and disease of cervical spine. Thus, medical experts first select several diagnostic candidates, which are very similar to each other, from many diseases and then make a final diagnosis from those candidates.

This paper formalizes these procedures from the viewpoint of rough sets[5] and introduces a new approach to rule induction.

3 Rough Set Theory and Probabilistic Rules

In the following sections, we use the following notations introduced by Grzymala-Busse and Skowron[8], which are based on rough set theory[5]. 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, $IS = (U, A \cup \{d\})$. 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, $f = [location = occular]$ and $f_A = \{1, 5, 6, 7\}$. As an example of a conjunctive formula, $g = [location = occular] \wedge [nausea = no]$ is a descriptor of U and g_A is equal to $\{1, 5\}$.

It is also notable that d can be treated as a formula (or an attribute-value pair) because $BsubseteqA$ is extended into $BsubseteqA \cup d$ and d has the same nature as an atttribute $a \in A$: that is, since d is of the form $[d = class_i]$, $d_A = \{s \in U | d(s) = class_i\}$. For simplicity, d_A is denoted by D in subsequent sections.

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

Table 1. A small example of a database

No.	loc	nat	his	prod	jolt	nau	M1	M2	class
1	occular	per	per	0	0	0	1	1	m.c.h.
2	whole	per	per	0	0	0	1	1	m.c.h.
3	lateral	thr	par	0	1	1	0	0	common.
4	lateral	thr	par	1	1	1	0	0	classic.
5	occular	per	per	0	0	0	1	1	psycho.
6	occular	per	subacute	0	1	1	0	0	i.m.l.
7	occular	per	acute	0	1	1	0	0	psycho.
8	whole	per	chronic	0	0	0	0	0	i.m.l.
9	lateral	thr	per	0	1	1	0	0	common.
10	whole	per	per	0	0	0	1	1	m.c.h.

Definition. loc: location, nat: nature, his:history,
Definition. prod: prodrome, nau: nausea, jolt: Jolt headache,
M1, M2: tenderness of M1 and M2, 1: Yes, 0: No, per: persistent,
thr: throbbing, par: paroxysmal, m.c.h.: muscle contraction headache,
psycho.: psychogenic pain, i.m.l.: intracranial mass lesion, common.:
common migraine, and classic.: classical migraine.

Definition 1.
Let R and D denote a formula in $F(B, V)$ and a meaning of a decision d. Classification accuracy and coverage(true positive rate) for $R \rightarrow d$ is defined as:

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

where $|S|$, $\alpha_R(D)$, $\kappa_R(D)$ 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), respectively.

In the above example, when R and D are set to $[nau = 1]$ and $[class = common]$, $\alpha_R(D) = 2/5 = 0.4$ 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 $R \leftrightarrow D$.

Finally, we define partial order of equivalence as follows:

Definition 2. *Let R_i and R_j be the formulae in $F(B, V)$ and let $A(R_i)$ denote a set whose elements are the attribute-value pairs of the form $[a, v]$ included in R_i. If $A(R_i) \subseteq A(R_j)$, then we represent this relation as:*

$$R_i \preceq R_j.$$

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

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

where δ_α and δ_κ denote given thresholds for accuracy and coverage, respectively.

4 Characterization Sets

4.1 Characterization Sets

In order to model medical reasoning, a statistical measure, coverage defined ins Section 2 plays an important role in modeling, which is equivalent to 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,

Definition 3. *Let R denote a formula in $F(B, V)$. Characterization sets of a target concept (D) is defined as:*

$$L_{\delta_\kappa}(D) = \{R|\kappa_R(D) \geq \delta_\kappa\}$$

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

$$\begin{array}{ll} \text{Independent type:} & L_{\delta_\kappa}(D_i) \cap L_{\delta_\kappa}(D_j) = \phi, \\ \text{Boundary type:} & L_{\delta_\kappa}(D_i) \cap L_{\delta_\kappa}(D_j) \neq \phi, \text{ and} \\ \text{Subcategory type:} & L_{\delta_\kappa}(D_i) \subseteq L_{\delta_\kappa}(D_j). \end{array}$$

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

Tsumoto focuses on the subcategory type in [10] because D_i and D_j cannot be differentiated by using the characterization set of D_j, which suggests that D_i is a generalized disease of D_j. Then, Tsumoto generalizes the above rule induction method into the overlapped type, considering rough inclusion[11]. However, both studies assumes two-level diagnostic steps: focusing mechanism and differential diagnosis, where the former selects diagnostic candidates from the whole classes and the latter makes a differential diagnosis between the focused classes.

The proposed method below extends these methods into multi-level steps.

4.2 Characteristics

We consider the special case of characterization sets in which the thresholds of coverage is equal to 1.0. That is,

$$L_{1.0}(D) = \{R_i|\kappa_{R_i}(D) = 1.0\}$$

Then, we have several interesting characteristics.

Theorem 1. *Let R_i and R_j two formulae in $L_{1.0}(D)$ such that $R_i \preceq R_j$. Then, $\alpha_{R_i} \leq \alpha_{R_j}$.*

Thus, when we collect the formulae whose values of coverage are equal to 1.0, the sequence of conjunctive formulae corresponds to the sequence of increasing chain of accuracies.

For example, $[nat = per]$ and $[his = per]$ are elements of $L_{1.0}(m.c.h.)$ and those accuracies are: 3/7 and 3/5. Then, since the meaning of $([loc = occular] \vee [loc = whole]) \wedge [his = per]$ is equal to $[1, 2, 5, 10]$, the accuracy of $[nat = per] \wedge [his = per]$ is 3/4.

Since $\kappa_R(D) = 1.0$ means that the meaning of R covers all the samples of D, its complement $U - R_A$, that is, $\neg R$ do not cover any samples of D. Especially, when R consists of the formulae with the same attributes, it can be viewed as the generation of the coarsest partitions. Thus,

Theorem 2. *Let R be a formula in $L_{1.0}(D)$ such that $R = \vee_j [a_i = v_j]$. Then, R and $\neg R$ gives the coarsest partition for a_i, whose R includes D.*

From the propositions 1 and 2, the next theorem holds.

Theorem 3. *Let A consist of $\{a_1, a_2, \cdots, a_n\}$ and R_i be a formula in $L_{1.0}(D)$ such that $R_i = \vee_j [a_i = v_j]$. Then, a sequence of a conjunctive formula $F(k) = \wedge_{i=1}^{k} R_i$ gives a sequence which increases the accuracy.* □

5 Rule Induction with Grouping

As discussed in Section 2, When the coverage of R for a target concept D is equal to 1.0, R is a necessity condition of D. That is, a proposition $D \rightarrow R$ holds and its contrapositive $\neg R \rightarrow \neg D$ holds. Thus, if R is not observed, D cannot be a candidate of a target concept. Thus, if two target concepts have a common formula R whose coverage is equal to 1.0, then $\neg R$ supports the negation of two concepts, which means these two concepts belong to the same group. Furthermore, if two target concepts have similar formulae $R_i, R_j \in L_{1.0}(D)$, they are very close to each other with respect to the negation of two concepts. In this case, the attribute-value pairs in the intersection of $L_{1.0}(D_i)$ and $L_{1.0}(D_j)$ give a characterization set of the concept that unifies D_i and D_j, D_k. Then, compared with D_k and other target concepts, classification rules for D_k can be obtained. When we have a sequence of grouping, classification rules for a given target concepts are defined as a sequence of subrules. From these ideas, a rule induction algorithm with grouping target concepts can be described as Figure 1. This algorithm first calculates $L_{1.0}(D_i)$ for $\{D_1, D_2, \cdots, D_k\}$. Second, from the list of characterization sets, it calculates the intersection between $L_{1.0}(D_i)$ and $L_{1.0}(D_j)$ and stores it into L_{id}. Third, the procedure calculates the similarity (matching number)of the intersections and sorts L_{id} with respect of the similarities. Fourth, the algorithm chooses one intersection $(D_i \cap D_j)$ with maximum similarity (highest matching number) and group D_i and D_j into a concept DD_i. These procedures will be continued until all the grouping is considered (Fig. 2). Finally, rules for generated group and diseases are induced by using a rule induction algorithm shown in Fig. 3.

```
procedure Total Process;
  var inputs
    L_D : List; /* A list of Target Concepts */
  begin
    Calculate a set of characterization set L_c;
      Calculate a set of intersection L_id;
        Calculate a list of similarity measures L_s;
          Calculate a list of grouping L_g; (Fig. 2)
        Induce a set of rules for L_g: L_r; (Fig. 3)
      Combine Rules in L_r for each D_i;
  end {Total Process}
```

Fig. 1. An Algorithm for Total Process

```
procedure Grouping ;
  var inputs
    L_c : List;
    /* A list of Characterization Sets */
    L_id : List;
    /* A list of Intersection */
    L_s : List;
    /* A list of Similarity */
  var outputs
    L_gr : List;
    /* A list of Grouping */
  var
    k : integer;        L_g, L_gr : List;
  begin
    L_g := {} ;
    k := n
    /* n: A number of Target Concepts*/
    Sort L_s with respect to similarities;
      Take a set of (D_i, D_j), L_max
        with maximum similarity values;
      k:= k+1;
      forall (D_i, D_j) ∈ L_max do
        begin
          Group D_i and D_j into D_k;
            L_c := L_c − {(D_i, L_1.0(D_i)};
              L_c := L_c − {(D_j, L_1.0(D_j)};
              L_c := L_c + {(D_k, L_1.0(D_k)};
              Update L_id for DD_k;
              Update L_s;
            L_gr := (
              Grouping for L_c, L_id, and L_s) ;
          L_g := L_g + {{(D_k, D_i, D_j), L_g}};
        end
    return L_g;
  end {Grouping}
```

Fig. 2. An Algorithm for Grouping

```
procedure RuleInduction ;
  var inputs
    L_c : List;
    /* A list of Characterization Sets */
    L_id : List; /* A list of Intersection */
    L_g : List; /* A list of grouping*/
    /* {{(D_n+1,D_i,D_j),{(DD_n+2,.)...}}} */
    /* n: A number of Target Concepts */
  var
    Q, L_r : List;
  begin
    Q := L_g; L_r := {};
    if (Q ≠ ∅) then do
      begin
        Q := Q − first(Q);
        L_r := Rule Induction (L_c, L_id, Q);
      end
    (DD_k, D_i, D_j) := first(Q);
    if (D_i ∈ L_c and D_j ∈ L_c) then do
      begin
        Induce a Rule r which discriminate
        between D_i and D_j;
        r = {R_i → D_i, R_j → D_j};
      end
    else do
      begin
        Search for L_1.0(D_i) from L_c;
        Search for L_1.0(D_j) from L_c;
        if (i < j) then do
          begin
            r(D_i) := ∨_{R_l ∈ L_1.0(D_j)} ¬R_l → ¬D_j;
            r(D_j) := ∧_{R_l ∈ L_1.0(D_j)} R_l → D_j;
          end
        r := {r(D_i), r(D_j)};
      end
    return L_r := {r, L_r} ;
  end {Rule Induction}
```

Fig. 3. An Algorithm for Rule Induction

6 Example

Let us consider Table 1 as an example for rule induction. For a similarity function, we use a matching number[3] which is defined as the cardinality of the intersection of two the sets. Also, since Table 1 has five classes, k is set to 6.

6.1 Grouping

From this table, the characterization set for each concept is obtained as shown in Fig 4. Then, the intersection between two target concepts are calculated. Since *common* and *classic* have the maximum matching number, these two classes are grouped into one category, D_6. Then, teh characterization of D_6 is obtained as :

$D_6 = \{[loc = lateral], [nat = thr], [jolt = 1], [nau = 1], [M1 = 0], [M2 = 0]$ from Fig 5.

In the second iteration, the intersection of D_1 and others is considered as shown in Fig 6. From this matrix, we have two possibilities of grouping: one is to group *m.c.h.* and *i.m.l.* That is, these two diseases are grouped into D_7: $D_7 = \{([loc = occular] \vee [loc = whole]), [nat = per], [prod = 0]\}$ The other one is to group D_1 and *i.m.l.*, where $D_7 = \{[jolt = 1], [M1 = 0], [M2 = 0]\}$.

In the third iteration of the former case(3_a), the intersection is calculated as Fig 7 and D_2 and psycho are grouped into D_3: $D_{3a} =$ { [nat=per], [prod=0] } In the latter case(3_b), it is calculated as Fig 8 and *m.c.h.* and *psycho* are grouped into D_8: $D_{8a} =$ { [nat=per], [prod=0] }. Fig 9 and 10 depicts the two results of grouping like a dendrogram in clustering analysis[3].

$$
\begin{aligned}
L_{1.0}(m.c.h.) = \quad & \{([loc = occular] \vee [loc = whole]), [nat = per], [his = per], \\
& [prod = 0], [jolt = 0], [nau = 0], [M1 = 1], [M2 = 1]\} \\
L_{1.0}(common) = \ & \{[loc = lateral], [nat = thr], ([his = per] \vee [his = par]), [prod = 0], \\
& [jolt = 1], [nau = 1], [M1 = 0], [M2 = 0]\} \\
L_{1.0}(classic) = \quad & \{[loc = lateral], [nat = thr], [his = par], [prod = 1], \\
& [jolt = 1], [nau = 1], [M1 = 0], [M2 = 0]\} \\
L_{1.0}(i.m.l.) = \quad & \{([loc = occular] \vee [loc = whole]), [nat = per], \\
& ([his = subacute] \vee [his = chronic]), [prod = 0], \\
& [jolt = 1], [M1 = 1], [M2 = 1]\} \\
L_{1.0}(psycho) = \quad & \{[loc = occular], [nat = per], ([his = per] \vee [his = acute]), \\
& [prod = 0]\}
\end{aligned}
$$

Fig. 4. Characterization Sets for Table 1

6.2 Rule Induction

Due to the limitation of space, we focus on rule induction based on the first model. Figure 9 shows one candidate of the differential diagnosis. For the differential diagnosis of *common*. First, this model discriminate between D_6(*common*

	m.c.h.	common	classic	i.m.l.	psycho
m.c.h.	–	{[prod=0]}	∅	{([loc=occular]∨[loc=whole]), {[nat=per],[prod=0]} {[nat=per],[prod=0]}	
common	–	–	{[loc=lateral], [nat=thr],[jolt=1], [nau=1], [M1=0],[M2=0]}	{[prod=0],[jolt=1], [M1=0], [M2=0] }	{[prod=0]}
classic	–	–	–	{[jolt=1],[M1=0],[M2=0]}	{ }
i.m.l.	–	– [prod=0]}	–	–	{[nat=per],

Fig. 5. Intersection of Two Characterization Sets (Step 2)

	m.c.h.	D_6	i.m.l.	psycho
m.c.h.	–	{}	{([loc=occular]∨[loc=whole]), {[nat=per],[prod=0]} {[nat=per],[prod=0]}	{[nat=per],[prod=0]}
D_6	–	–	{[jolt=1], [M1=0], [M2=0]}	{ }
i.m.l.	–	–	–	{[nat=per],[prod=0]}

Fig. 6. Intersection of Two Characterization Sets after the first Grouping (Step 3)

	D_6	D_7	psycho
D_6	–	{}	{ }
D_7	–	–	{[nat=per],[prod=0]}

Fig. 7. Intersection of Two Characterization Sets after the first Grouping (1) (Step 4a)

	m.c.h.	D_7	psycho
m.c.h.	–	{}	{[nat=per],[prod=0] }
D_7	–	{}	{ }

Fig. 8. Intersection of Two Characterization Sets after the first Grouping (2) (Step 4b)

and *classic*) and D_8 (*m.c.h.*, *i.m.l.* and *psycho*). Then, *common* and *classic* within D_6 are differentiated. Thus, a classification rule for *common* is composed of two subrules: (discrimination between D_6 and D_8) and (discrimination within D_6). On the other hand, a classification rule for *m.c.h.* is composed of three subrules: (discrimination between D_6 and D_8), (discrimination between D_7 and psycho) and (discrimination within D_7).

Let us consider the first case. The first part can be obtained by the intersection in Figure 7. That is, $D_8 \rightarrow [nat = per] \wedge [prod = 0]$; $\neg[nat = per] \vee \neg[prod = 0] \rightarrow \neg D_8$. Then, since from Figure 4, the difference set between $L_{1.0}(common)$ and $L_{1.0}(classic)$ is $\{[prod = 1]\}$, for a classification rule for *common* within D_7 is: $[prod = 0] \rightarrow common$.

Combining these two parts, the classification rule for *common* is: $(\neg[nat = per] \vee \neg[prod = 0]) \wedge [prod = 0] \rightarrow common$. After its simplification, the rule is:

$$\neg[nat = per] \rightarrow \neg common,$$

whose accuracy is equal to 2/3. In the same way, the rule for *classic* is obtained as:

$$\neg[nat = per] \wedge [prod = 1] \rightarrow classic.$$

7 Experimental Results

The above rule induction algorithm was implemented in PRIMEROSE4.5 (Probabilistic Rule Induction Method based on Rough Sets Ver 5.0), and was applied

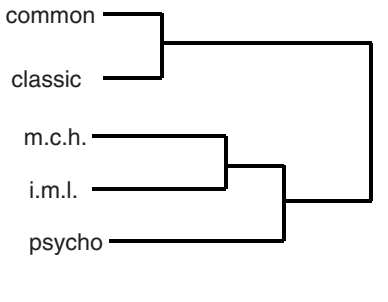

Fig. 9. Grouping by Characterization Sets (First Model)

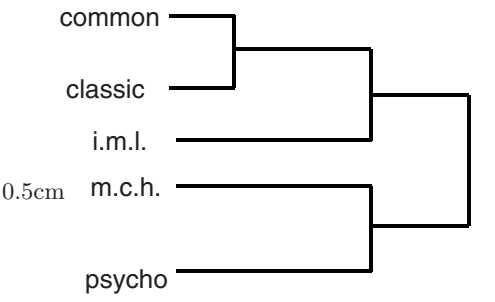

Fig. 10. Grouping by Characterization Sets (Second Model)

to databases on differential diagnosis of headache, meningitis and cerebrovascular diseases (CVD), whose precise information is given in Table 2. In these experiments, δ_α and δ_κ were set to 0.75 and 0.5, respectively. Also, the threshold for grouping is set to 0.8 [1]. This system was compared with PRIMEROSE4.5[11], PRIMEROSE[9] C4.5[6], CN2[2], AQ15[4] with respect to the following points: length of rules, similarities between induced rules and expert's rules and performance of rules.

In this experiment, the 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[3]. Concerning the performance of rules, ten-fold cross-validation was applied to estimate classification accuracy.

Table 2. Information about Databases

Domain	Samples	Classes	Attributes
Headache	52119	45	147
CVD	7620	22	285
Meningitis	141	4	41

Table 3 shows the experimental results, which suggest that PRIMEROSE5 outperforms PRIMEROSE4.5 (two-level) and the other four rule induction methods and induces rules very similar to medical experts' ones.

8 Discussion

The readers may wonder why lengthy rules perform better than short rules since lengthy rules suffer from overfitting to a given data. One reason is that a decision

[1] These values are given by medical experts as good thresholds for rules in these three domains.

Table 3. Experimental Results

Method	Length	Similarity	Accuracy
	Headache		
PRIMEROSE5.0	8.8 ± 0.27	0.95 ± 0.08	$95.2 \pm 2.7\%$
PRIMEROSE4.5	7.3 ± 0.35	0.74 ± 0.05	$88.3 \pm 3.6\%$
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	$86.2 \pm 2.9\%$
	Meningitis		
PRIMEROSE5.0	2.6 ± 0.19	0.91 ± 0.08	$82.0 \pm 3.7\%$
PRIMEROSE4.5	2.8 ± 0.45	0.72 ± 0.25	$81.1 \pm 2.5\%$
Experts	3.1 ± 0.32	1.00 ± 0.00	$85.0 \pm 1.9\%$
PRIMEROSE	1.8 ± 0.45	0.64 ± 0.25	$72.1 \pm 2.5\%$
C4.5	1.9 ± 0.47	0.63 ± 0.20	$73.8 \pm 2.3\%$
CN2	1.8 ± 0.54	0.62 ± 0.36	$75.0 \pm 3.5\%$
AQ15	1.7 ± 0.44	0.65 ± 0.19	$74.7 \pm 3.3\%$
	CVD		
PRIMEROSE5.0	7.6 ± 0.37	0.89 ± 0.05	$74.3 \pm 3.2\%$
PRIMEROSE4.5	5.9 ± 0.35	0.71 ± 0.05	$72.3 \pm 3.1\%$
Experts	8.5 ± 0.43	1.00 ± 0.00	$82.9 \pm 2.8\%$
PRIMEROSE	4.3 ± 0.35	0.69 ± 0.05	$74.3 \pm 3.1\%$
C4.5	4.0 ± 0.49	0.65 ± 0.09	$69.7 \pm 2.9\%$
CN2	4.1 ± 0.44	0.64 ± 0.10	$68.7 \pm 3.4\%$
AQ15	4.2 ± 0.47	0.68 ± 0.08	$68.9 \pm 2.3\%$

attribute gives a partition of datasets: since the number of given classes are 4 to 45, some classes have very low support due to the prevalence of the corresponding diseases. Thus, the disease with the low frequency may not have short-length rules by using the conventional methods. However, since our method is not based on accuracy, but on coverage, we can support the disease of frequency. Another reason is that this method reflects the reasoning style of domain experts. One of the most important features 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 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.

9 Conclusion

In this paper, the characteristics of experts' rules are closely examined, whose empirical results suggest that grouping of diseases ais 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 for grouping of diseases by using this measure. The above example shows 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. This research is a preliminary study on a rule induction method with grouping and it will be a basis for a future work to compare the proposed method with other rule induction methods by using real-world datasets.

Acknowledgments

The author thanks for three reviewers' insightful comments. This work was supported by the Grant-in-Aid for Scientific Research (13131208) on Priority Areas (No.759) "Implementation of Active Mining in the Era of Information Flood" by the Ministry of Education, Science, Culture, Sports, Science and Technology of Japan.

References

1. Aha, D. W., Kibler, D., and Albert, M. K., Instance-based learning algorithm. *Machine Learning*, 6, 37-66, 1991.
2. Clark, P. and Niblett, T., The CN2 Induction Algorithm. *Machine Learning*, 3, 261-283, 1989.
3. Everitt, B. S., *Cluster Analysis*, 3rd Edition, John Wiley & Son, London, 1996.
4. 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.
5. Pawlak, Z., *Rough Sets*. Kluwer Academic Publishers, Dordrecht, 1991.
6. Quinlan, J.R., *C4.5 - Programs for Machine Learning*, Morgan Kaufmann, Palo Alto, 1993.
7. *Readings in Machine Learning*, (Shavlik, J. W. and Dietterich, T.G., eds.) Morgan Kaufmann, Palo Alto, 1990.
8. 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.
9. Tsumoto, S., Automated Induction of Medical Expert System Rules from Clinical Databases based on Rough Set Theory. *Information Sciences* **112**, 67-84, 1998.
10. Tsumoto, S., Extraction of Experts' Decision Rules from Clinical Databases using Rough Set Model *Intelligent Data Analysis*, 2(3), 1998.
11. Tsumoto,S. Extraction of Hierarchical Decision Rules from Clinical Databases using Rough Sets. *Information Sciences*, 2003 (in print)