Mining Diagnostic Taxonomy for Multi-Stage Medical Diagnosis

Shusaku Tsumoto

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

Abstract. Experts' reasoning selects the final diagnosis from many candidates by using hierarchical differential diagnosis. In other words, candidates gives a sophisticated hiearchical taxonomy, usually described as a tree. In this paper, the characteristics of experts' rules are closely examined from the viewpoint of hierarchical decision steps and and a new approach to rule mining with extraction of diagnostic taxonomy from medical datasets is introduced. The key elements of this approach are calculation of the characterization set of each decision attribute (a given class) and one of the similarities between characterization sets. From the relations between similarities, tree-based taxonomy is obtained, which includes enough information for hierarchical diagnosis. The proposed method was evaluated on three medical datasets, the experimental results of which show that induced rules correctly represent experts' decision processes.

Keywords Rough sets, data mining, taxonomy, granular computing.

1 Introduction

Rule mining has been applied to many domains. However, empirical results show that the interpretation of extracted rules requires deep understanding for applied domains. One of its reasons is that conventional rule induction methods such as C4.5[6] cannot reflect the type of experts' reasoning. For example, rule induction methods such as AQ15[4], PRIMEROSE[9] induce the following common rule for muscle contraction headache from databases on differential diagnosis of headache:

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

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

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

where [Tenderness of B1 = *no*] and [Tenderness of C1 = *no*] are added. It is notable that these observation can be found in several medical domains[9].

One of the main reasons why the rules obtained from the dataset are shorter is that these patterns are generated only by a simple criteria, such as high accuracy or high information gain. The comparative studies[9–11] suggest that experts should acquire rules not only by a single 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*]
$\wedge$([Tenderness of M0 = *yes*] $\vee$[Tenderness of M1 = *yes*] $\vee$[Tenderness of M2 = *yes*])
$\wedge$([Tenderness of B1 = *yes*] $\vee$[Tenderness of B2 = *yes*] $\vee$[Tenderness of B3 = *yes*]
  $\vee$[Tenderness of C1 = *yes*] $\vee$[Tenderness of C2 = *yes*] $\vee$[Tenderness of C3 = *yes*]
  $\vee$[Tenderness of C4 = *yes*])
    $\rightarrow$ disease of cervical spine

The differences between these two rules are attribute-value pairs, from tenderness of B1 to C4. Thus, these two rules are composed of the following three blocks:

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

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*] $\vee$ [Tenderness of $M1$ = *yes*] $\vee$ [Tenderness of M2 = *yes*], and A_3 = [Tenderness of C1 = *no*] $\wedge$ [Tenderness of C2 = *no*] $\wedge$ [Tenderness of C3 = *no*] $\wedge$ [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 the differential diagnosis between headaches caused by neck muscles and ones by 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.

In this paper, the characteristics of experts' rules are closely examined from the viewpoint of hierarchical decision steps. Then, extraction of diagnostic taxonomy from medical datasets is introduced, which consists of the following three procedures. First, the characterization set of each attribute-value pair for a decision attribute(a given class) is extracted from databases. Then, similarities between the characterization sets are calculated. Finally, the concept hierarchy for given classes is calculated from the similarity values.

The paper is organized as follows. Section 2 and 3 introduces rough sets and a characterization set. Section 4 gives an algorithm for extraction of diagnostic taxonomy. Section 5 shows an illustrative example. Section 6 gives how rules are induced after grouping. Finally, Section 7 concludes this paper. The proposed method was evaluated on medical databases, the experimental results of which show that induced rules correctly represent experts' decision processes.

2 Rough Set Theory: Preliminaries

In the following sections, we use the following notations introduced by Grzymala-Busse and Skowron[8], which are based on rough set theory[5].

Let U denote a nonempty, finite set called the universe and A denote a nonempty, finite set of conditional attributes, i.e., $a : U \rightarrow V_a$ for $a \in A$, where V_a is called the domain of a, respectively. For A, V_A denotes a set of the domain of attributes. Then, a decision table is defined as an information system, $A = (U, A \cup \{d\})$, where $\{d\}$ denotes a decision attribute (a set of given classes).

The atomic formulae over $B \subseteq A \cup \{d\}$ and V_B are expressions of the form $[a = v]$, called descriptors over B, where $a \in B$ and $v \in V_a$. The set $F(B, V_B)$ of formulas over B is the least set containing all atomic formulas over B and closed with respect to disjunction, conjunction and negation. For each $f \in F(B, V_B)$, 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) $(f \wedge g)_A = f_A \cap g_A$; $(f \vee g)_A = f_A \vee g_A$; $(\neg f)_A = U - f_a$

By the use of the framework above, classification accuracy and coverage are defined as follows.

Definition 1.
Let R denote a formula in $F(B, V_B)$ and D a set of objects which belong to a decision attribute d. Classification accuracy and coverage(true positive rate) for $R \rightarrow d$ is defined as:

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

where $|S|$, $\alpha_R(D)$, $\kappa_R(D)$ and P(S) denote the cardinality of a set S, a classification accuracy and coverage of R as to classification of D, and probability of S, respectively.

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.

Also, we define partial order of equivalence as follows:

Definition 2. *Let R_i and R_j be the formulae in $F(B, V_B)$ 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: $R_i \preceq R_j$.*

Finally, according to the above definitions, probabilistic rules with high accuracy and coverage are defined as:

$$R \overset{\alpha,\kappa}{\rightarrow} d \; s.t. \; R = \wedge_i [a_i = v_k], \; \alpha_R(D) \geq \delta_\alpha \text{ and } \kappa_R(D) \geq \delta_\kappa,$$

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

3 Characterization Sets

3.1 Characterization Sets

In order to model medical reasoning, a statistical measure, coverage plays an important role in modeling. Let us define a characterization set of D, denoted by $L_{\delta_\kappa}(D)$ 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 $F(B, V_B)$. Characterization sets of a decision attribute (D) is defined as:*

$$L_{\delta_\kappa}(D) = \{R | R_i = \wedge_i (\vee_j [a_i = v_j]) \text{ and } \kappa_R(D) \geq \delta_\kappa\},$$

Then, three types of relations between characterization sets can be defined as follows: (1) Independent type: $L_{\delta_\kappa}(D_i) \cap L_{\delta_\kappa}(D_j) = \phi$, (2) Overlapped type: $L_{\delta_\kappa}(D_i) \cap L_{\delta_\kappa}(D_j) \neq \phi$, and (3) Subcategory 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, 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. In this paper, we consider the special case of characterization sets in which each formulae is given as a conjunctive normal formula and the thresholds of coverage is equal to 1.0: $L_{1.0}(D) = \{R_i | R_i = \wedge_i (\vee_j [a_i = v_j]), \quad \kappa_{R_i(D)} = 1.0\}$ It is notable that this set has several interesting characteristics.

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

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.*

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.*

4 Rule Induction with Diagnostic Taxonomy

4.1 Intuitive Ideas

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. It means that if R is not observed, D cannot be a candidate of a decision class. If two decision classes have a common formula R whose coverage is equal to 1.0, then $\neg R$ supports the negation of two classes, 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 classes. 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 generalized decision class that unifies D_i and D_j, DD_k. Then, compared with DD_k and other target concepts, classification rules for DD_k can be obtained. When we have a sequence of grouping, classification rules for a given decision classes are defined as a sequence of subrules. From these ideas, a rule induction algorithm with grouping target concepts can be described as a combination of grouping (Figure 1) and rule induction (Figure 2). First, 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. The first to fourth steps are described as Figure 1. Finally, rules for each decision class, including grouped ones, are induced. For given decision classes, rules are composed of rules for the upper-level and rules specific to the corresponding given class shown in Figure 2.

4.2 Similarity

To measure the similarity between two characterization sets, we can apply several indices of two-way contigency tables. Table 1 gives a contingency table for two rules, $L_{1.0}(D_i)$ and $L_{1.0}(D_j)$. The first cell a (the intersection of the first row and column) shows the number of matched attribute-value pairs. From this table, several kinds of similarity measures can be defined. The best similarity measures in the statistical literature are four measures shown in Table 2[3, 2].

In this paper, we focus on the two similarity measures: one is Simpson's measure: $\frac{a}{min\{(a+b),(a+c)\}}$ and the other is Braun's measure: $\frac{a}{max\{(a+b),(a+c)\}}$.

As discussed in Subsection 4.2, a single-valued similarity becomes low when $L_{1.0}(D_i) \subset L_{1.0}(D_j)$ and $|L_{1.0}(D_i)| << |L_{1.0}(D_j)|$. For example, let us consider when $|L_{1.0}(D_i)| = 1$. Then, match number is equal to 1.0, which is the lowest value of this similarity. In the case of Jaccard's coefficient, the value is $1/1 + b$ or $1/1 + c$: the similarity is very small when $1 << b$ or $1 << c$. Thus, these

```
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
    $k$ : $integer$;      $L_g$ : $List$;
  begin
    $L_g := \{\}$ ; $L_{gr} := \{\}$;
    $k := n$ /* 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 $(D_i, D_j) \in L_{max}$ do
        begin
          $L_g := \{\}$;
          Group $D_i$ and $D_j$ into $DD_k$;
            $L_c := L_c - \{(D_i, L_{1.0}(D_i)\}$;
              $L_c := L_c - \{(D_j, L_{1.0}(D_j)\}$;
              $L_c := L_c + \{(D_k, L_{1.0}(D_k)\}$;
              Update $L_{id}$ for $DD_k$;
              Update $L_s$;
            $L_g :=$ (Outputs from $Grouping$ for $L_c$, $L_{id}$, and $L_s$) ;
          $L_{gr} := L_{gr} + \{\{(DD_k, D_i, D_j), L_g\}\}$;
        end
    return $L_{gr}$;
  end $\{Grouping\}$
```

Fig. 1. An Algorithm for Grouping

similarities do not reflect the subcategory type. Thus, we should check the difference between $a + b$ and $a + c$ to consider the subcategory type. One solution is to take an interval of maximum and minimum as a similarity, which we call an interval-valued similarity.

For this purpose, we combine Simpson and Braun similarities and define an interval-valued similarity: $\left[\frac{a}{max\{(a+b),(a+c)\}}, \frac{a}{min\{(a+b),(a+c)\}}\right]$ If the difference between two values is large, it would be better not to consider this similarity for grouping in the lower generalization level. For example, when $a + c = 1 (a = 1, c = 0)$, the above value will be: $\left[\frac{1}{1+b}, 1\right]$ If $b >> 1$, then this similarity should be kept as the final candidate for the grouping.

The disadvantage is that it is difficult to compare these interval values. In this paper, the maximum value of a given interval is taken as the representative of this similarity when the difference between min and max are not so large.

```
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 $(Q = \emptyset)$ then return $L_r = \{\}$;
    if $(Q \neq \emptyset)$ 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 $(D_i \in L_c$ and $D_j \in L_c)$ then do
      begin
        Induce a Rule $r$ which discriminate
        between $D_i$ and $D_j$;
        $r = \{R_i \rightarrow D_i, R_j \rightarrow D_j\}$;
      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
            $r(D_i) := \vee_{R_l \in L_{1.0}(D_j)} \neg R_l \rightarrow \neg D_j$;
            $r(D_j) := \wedge_{R_l \in L_{1.0}(D_j)} R_l \rightarrow D_j$;
          end
        $r := \{r(D_i), r(D_j)\}$;
      end
    return $L_r := \{r, L_r\}$ ;
  end {Rule Induction}
```

Fig. 2. An Algorithm for Rule Induction

If the maximum values are equal to the other, then the minimum value will be compared. If the minimum value is larger than the other, the larger one is selected.

Table 1. Contingency Table for Similarity

		$L_{1.0}(D_j)$		
		Observed	*Not Observed*	Total
$L_{1.0}(D_i)$	*Observed*	a	b	$a+b$
	Not observed	c	d	$c+d$
	Total	$a+c$	$b+d$	$a+b+c+d$

Table 2. A List of Similarity Measures

(1) Matching Number	a
(2) Jaccard's coefficient	$a/(a+b+c)$
(3) χ^2-statistic	$N(ad-bc)^2/M$
(4) point correlation coefficient	$(ad-bc)/\sqrt{M}$
(5) Kulczynski	$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$
(6) Ochiai	$\frac{a}{\sqrt{(a+b)(a+c)}}$
(7) Simpson	$\frac{a}{min\{(a+b),(a+c)\}}$
(8) Braun	$\frac{a}{max\{(a+b),(a+c)\}}$

$N=a+b+c+d$, $M=(a+b)(b+c)(c+d)(d+a)$

5 Example

Let us consider Table 3 as an example for rule induction. For a similarity function, we use the interval similarity defined in Section 4.2. Since Table 3 has five classes in the decision attribute, an index for grouped concepts, k is set to 6. For extraction of taxonomy, the interval-valued similarity is applied.

5.1 Grouping

From this table, the characterization set for each concept is obtained as shown in Fig 3. Then, the intersection between two target concepts are calculated. In the first level, the similarity matrix is generated as shown in Fig. 4.

Since *common* and *classic* have the maximum similarity, these two classes are grouped into one category, D_6. Then, the characterization of D_6 is obtained as : $D_6 = \{[loc = lateral], [nat = thr], [jolt = 1], [nau = 1], [M1 = 0], [M2 = 0]\}$. In the second iteration, the intersection of D_6 and others is considered and the similarity matrix is obtained: as shown in Fig 5. From this matrix, we have to compare three candidates: [2/8,2/4], [3/7,3/6] and [2/7,2/4]. From the minimum values, the middle one: D_6 and *i.m.l.* is selected as the second grouping. Thus, $D_7 = \{[jolt = 1], [M1 = 0], [M2 = 0]\}$. In the third iteration, the intersection matrix is calculated as Fig 6 and *m.c.h.* and *psycho* are grouped into D_8: $D_8 =$ { [nat=per], [prod=0] }. Finally, the dendrogram is given as Fig. 7.

Table 3. 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.

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

Fig. 3. Characterization Sets for Table 3

	m.c.h.	common	classic	i.m.l.	psycho
m.c.h.	–	[1/8,1/8]	[0,0]	[3/8,3/7]	[2/8,2/4]
common	–	–	[6/8,6/8]	[4/8, 4/7]	[1/7,1/4]
classic	–	–	–	[3/8, 3/7]	0
i.m.l.	–	–	–	–	[2/7, 2/4]

Fig. 4. Interval-valued Similarity of Two Characterization Sets (Step 2)

5.2 Rule Induction

The grouping obtained from the dataset shows the candidate of the differential diagnosis taxonomy with the given interval-valued similarity. For differential diagnosis, First, this model discriminate between D_7(*common*, *classic* and *i.m.l.*)

	m.c.h.	D_6	i.m.l.	psycho
m.c.h.	–	0	[3/8, 3/7]	[2/8,2/4]
D_6	–	–	[3/7,3/6]	0
i.m.l.	–	–	–	[2/7,2/4]

Fig. 5. Interval-valued Similarity of Two Characterization Sets after the first Grouping (Step 3)

	m.c.h.	D_7	psycho
m.c.h.	–	$[0, 0]$	[2/8,2/4]
D_7	–	$[0, 0]$	[0,0]

Fig. 6. Interval-valued Similarity of Two Characterization Sets after the second Grouping (Step 4)

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

The first part can be obtained by the intersection for Figure 6. That is,

$$D_8 \rightarrow [nat = per] \wedge [prod = 0]$$

$$\neg[nat = per] \vee \neg[prod = 0] \rightarrow \neg D_8.$$

Then, the second part can be obtained by the intersection for Figure 5. That is,

$$\begin{aligned}\neg([loc = occular] \vee [loc = whole]) \vee \neg[nat = per] \\ \vee \neg([his = subacute] \vee [his = chronic]) \\ \vee \neg[prod = 0] \rightarrow \neg i.m.l.\end{aligned}$$

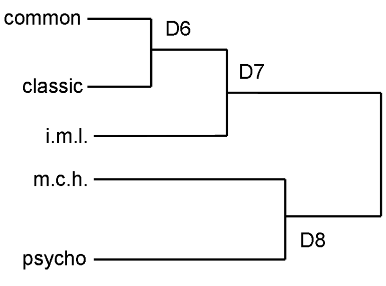

Fig. 7. Grouping by Characterization Sets

Finally, the third part of the rule can be obtained by the difference set between $L_{1.0}(common)$ and $L_{1.0}(classic) = \{[prod = 1]\}$.

$$[prod = 0] \rightarrow common.$$

Combining these three parts, the classification rule for *common* is

$$\begin{aligned}
&(\neg[nat = per] \vee \neg[prod = 0]) \\
&\qquad \wedge (\neg([loc = occular] \vee [loc = whole]) \vee \neg[nat = per] \\
&\qquad \vee \neg([his = subacute] \vee [his = chronic]) \vee \neg[prod = 0]) \\
&\qquad\qquad \wedge [prod = 0] \rightarrow common.
\end{aligned}$$

After its simplification, the rule is transformed into:

$$\begin{aligned}
&[nat = thr] \wedge ([loc = lateral] \vee \neg([his = subacute] \vee [his = chronic])) \\
&\qquad\qquad \wedge [prod = 0] \rightarrow common.
\end{aligned}$$

whose accuracy is equal to 2/3.

6 Experimental Results

The above rule induction algorithm was implemented in PRIMEROSE5.0 (Probabilistic Rule Induction Method based on Rough Sets Ver 5.0), and was applied to databases on differential diagnosis of headache, meningitis and cerebrovascular diseases (CVD), whose precise information is given in Table 4. 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[1], 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 for comparison[3]. Concerning the performance of rules, ten-fold cross-validation was applied to estimate classification accuracy.

Table 5 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.

7 Discussion

7.1 Focusing Mechanism

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 4. Information about Databases

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

Table 5. 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 with low 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 candidates, 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.

7.2 Sensitivity to Similarity

The problem with this approach is that several taxonomy trees are obtained when a single-valued similarity is adopted. If Simpson similarity is selected for grouping, two other models are acquired from the small dataset (Fig. 8,9). Although the model shown in Fig. 8 is topologically identical to Fig. 7, the grouping order is different. Thus, when the above rule induction method is applied, rules induced by this model may be different from the above rules. The other model is totally different from those two models, so the obtained rule will be different from the rule in Section 5.

Moreover, if the matching number is selected for grouping, the other model is acquired (Fig. 10).

The selection of the interval-valued similarity is a solution to this problem. However, since this choice may not prevent the multiple model generation in general, it will be our future work to introduce a preference criteria for model selection.

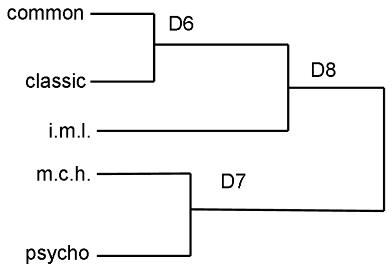

Fig. 8. The Second Grouping by Simpson Similarity

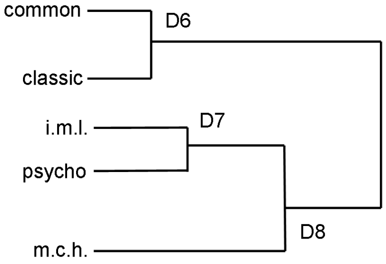

Fig. 9. The Third Grouping by Simpson Similarity

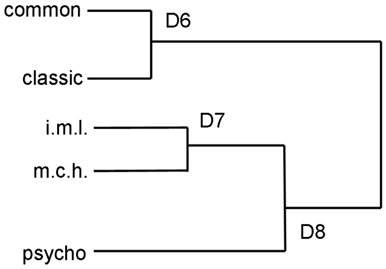

Fig. 10. The Second Grouping by Matching Number

8 Conclusion

In this paper, the characteristics of experts' rules are closely examined, whose empirical results suggest that grouping of diseases is 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, which consists of the following three procedures. First, the characterization set of each attribute-value pair for a decision class(a given class) is extracted from databases. Then, similarities between the characterization sets are calculated. Finally, the concept hierarchy for given classes is calculated from the similarity values. The proposed method was evaluated on three medical datasets, the experimental results of which show that induced rules correctly represent experts' decision processes.

Although the proposed method gives a good performance with diagnostic taxonomy, it is possible that the method outputs multiple models. This observation is dependent on the selection of the similarity measure. It will be our future work to solve this problem.

Acknowledgements

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. Clark, P. and Niblett, T., The CN2 Induction Algorithm. *Machine Learning*, 3, 261-283, 1989.
2. Cox, T. F. and Cox, M. A. A. *Multidimensional Scaling (Second Edition)*, Chapman & Hall/CRC, Boca Raton, 2000.
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. Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model. *Inforamtion Sciences* **162**, 65-80, 2004.