Mining Hierarchical Decision Rules from Clinical Databases Using Rough Sets and Medical Diagnostic Model

Shusaku Tsumoto

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

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 [8] cannot extract rules, which plausibly represent experts' decision processes [10]: the description length of induced rules is too short, compared with the experts' rules. For example, rule induction methods, including AQ15 [4] and PRIMEROSE [10], induce the following common rule for muscle contraction headache from databases on differential diagnosis of headache:

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

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

T. Elomaa et al. (Eds.): PKDD, LNAI 2431, pp. 423–435, 2002.

[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]
$\wedge$[Tenderness of C4 = no]
$\rightarrow$ muscle contraction headache

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

These results suggest that conventional rule induction methods do not reflect a mechanism of knowledge acquisition of medical experts.

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

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, 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, 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 can be simplified into the following form:

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

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. 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:

3 Rough Set Theory and Probabilistic Rules

3.1 Rough Set Notations

In the following sections, we use the following notations introduced by Grzymala-Busse and Skowron [9], 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, $A = (U, A \cup \{d\})$. 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 g_A is equal to $\{2, 5\}$.

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

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,

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:

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

where $|S|$, $\alpha_R(D)$, $\kappa_R(D)$ and P(S) denote the cardinality of a set S, a classification accuracy 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 $R \leftrightarrow D$.

3.2 Probabilistic Rules

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

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

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:

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

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

3.3 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:

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

All three definitions correspond to the negative region, boundary region, and positive region[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

$$\begin{array}{ll} L_{0.6}(m.c.h.) & = \{[age = 40-49], [location = whole], [nature = persistent], \\ & \quad [prodrome = no], [nausea = no], [M1 = yes]\}, \text{ and} \\ L_{0.6}(migraine) & = \{[age = 40-49], [nature = throbbing], \\ & \quad [nausea = yes], [M1 = no]\}, \end{array}$$

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

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

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.

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.

3.4 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 in their study on rough mereology [6]. 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 [12]. Let us illustrate how this measure is applied to hierarchical rule induction by using Table 1. When the threshold for the coverage is set to 0.6,

$$\mu(L_{0.6}(m.c.h.), L_{0.6}(migraine)) = \frac{|\{[age{=}40{-}49]\}|}{|\{[age{=}40{-}49],[location{=}whole],...\}|} = \frac{1}{6}$$
$$\mu(L_{0.6}(m.c.h.), L_{0.6}(psycho))$$
$$= \frac{|\{[location{=}whole],[nature{=}persistent],[prodrome{=}no],[M1{=}yes]\}|}{|\{[age{=}40{-}49],[location{=}whole],...\}|} = \frac{4}{6} = \frac{2}{3}$$
$$\mu(L_{0.6}(migraine), L_{0.6}(psycho))$$
$$= \frac{|\{[nausea{=}yes]\}|}{|\{[age{=}40{-}49],[nature{=}throbbing],...\}|} = \frac{1}{4}$$

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

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

4 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]

```
procedure Rule Induction (Total Process);
  var
    i : integer;    M, L, R : List;
     L_D : List; /* A list of all classes */
  begin
    Calculate α_R(D_i) and κ_R(D_i) for each elementary relation R and each class D_i;
    Make a list L(D_i) = {R|κ_R(D) = 1.0}) for each class D_i;
    while (L_D ≠ φ) do
      begin
         D_i := first(L_D); M := L_D − D_i;
        while (M ≠ φ) do
          begin
             D_j := first(M);
               if  (μ(L(D_j), L(D_i)) ≤ δ_μ) 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 };
```

Fig. 1. An algorithm for rule induction

[1] This method is an extension of PRIMEROSE4 reported in [11]. 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 Induction of Classification Rules;
  var
    i : integer;    M, L_i : List;
  begin
     L_1 := L_er; /* L_er: List of Elementary Relations */
    i := 1;    M := {};
    for i := 1 to n do       /* n: Total number of attributes */
      begin
        while ( L_i ≠ {} ) do
          begin
            Select one pair R = ∧[a_i = v_j] from L_i;
             L_i := L_i − {R};
            if   (α_R(D) ≥ δ_α)    and    (κ_R(D) ≥ δ_κ)
              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 };
```

Fig. 2. An algorithm for classification rules

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.

$$
\begin{array}{ll}
[M1 = yes] & \rightarrow D_1\ \alpha = 1.0, \kappa = 1.0, \text{ supported by } \{1,2,5,6\} \\
[prod = no] & \rightarrow D_1\ \alpha = 1.0, \kappa = 1.0, \text{ supported by } \{1,2,5,6\} \\
[nau = no] & \rightarrow D_1\ \alpha = 1.0, \kappa = 0.75, \text{supported by } \{1,2,5\} \\
[nat = per] & \rightarrow D_1\ \alpha = 1.0, \kappa = 0.75, \text{supported by } \{1,2,6\} \\
[loc = who] & \rightarrow D_1\ \alpha = 1.0, \kappa = 0.75, \text{supported by } \{2,5,6\} \\
[age = 50...59] & \rightarrow D_1\ \alpha = 1.0, \kappa = 0.5, \text{ supported by } \{2,6\}
\end{array}
$$

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.*.

```
procedure Rule Integration;
  var
    i : integer;    M, L2 : List; R(Di) : List; /* A list of rules for Di */
     LD : List; /* A list of all classes */
  begin
    while(LD ≠ φ) do
      begin
        Di := first(LD); M := L2(Di);
        Select one rule R' → D'i from R(L2(Di)).
        while (M ≠ φ) do
          begin
            Dj := first(M);
              Select one rule R → dj for Dj;
              Integrate two rules: R ∧ R' → dj.
            M := M − {Dj};
          end
        LD := LD − Di;
      end
  end {Rule Combination}
```

Fig. 3. An algorithm for rule integration

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

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

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

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

5 Experimental Results

The above rule induction algorithm was implemented in PRIMEROSE4.5 (Probabilistic Rule Induction Method based on Rough Sets Ver 4.5), and was applied

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

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.[2] This system was compared with PRIMEROSE4.0 [11], PRIMEROSE [10] C4.5 [7], 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, 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 4. Information about databases

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

Table 5 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.

6 Discussion: What Is Discovered?

Several interesting rules for migraine were found. Since migraine is a kind of vascular disease, the first part discriminates between migraine and other diseases. This part is obtained as :

$$[Nature : Persistent] \& \neg [History : acuteorparoxysmal]$$
$$\& [JoltHeadache : yes] \rightarrow \{commonmigraine, classicmigraine\}$$

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

Table 5. Experimental results

Method	Length	Similarity	Accuracy
	Headache		
PRIMEROSE4.5	8.8 ± 0.27	0.95 ± 0.08	$95.2 \pm 2.7\%$
PRIMEROSE4.0	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		
PRIMEROSE4.5	2.6 ± 0.19	0.91 ± 0.08	$82.0 \pm 3.7\%$
PRIMEROSE4.0	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		
PRIMEROSE4.5	7.6 ± 0.37	0.89 ± 0.05	$74.3 \pm 3.2\%$
PRIMEROSE4.0	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	$68.9 \pm 2.3\%$

which are reasonable for medical expert knowledge. Rather, medical experts pay attention to the corresponding parts and grouping of other diseases:

$$[Nature : Persistent]\&\neg[History : acuteorparoxysmal]$$
$$\&[JoltHeadache : yes] \rightarrow \{meningitis, Braintumor\},$$
$$[Nature : Persistent]\&\neg[History : acuteorparoxysmal]$$
$$\&[JoltHeadache : no] \rightarrow \{musculecontractionheadache\},$$

The former one is much more interesting and unexpected to medical experts, while the latter one is reasonable.

The second part discriminates between common migraine and classic migraine. These parts are obtained as :

$$[Age > 40]\&[Prodrome : no] \rightarrow CommonMigraineand$$
$$[Age < 20]\&[Prodrome : yes] \rightarrow ClassicMigraine,$$

where the attribute age is unexpected to medical experts. Migraine can be observed mainly by women, and it is observed that the frequency of headche decreases as women are getting older. Thus, the factor age support these experiences.

7 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 (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. 432
3. Everitt, B. S., *Cluster Analysis*, 3rd Edition, John Wiley & Son, London, 1996. 432
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. 423, 432
5. Pawlak, Z., *Rough Sets*. Kluwer Academic Publishers, Dordrecht, 1991. 425
6. Polkowski, L. and Skowron, A.: Rough mereology: a new paradigm for approximate reasoning. Intern. J. Approx. Reasoning **15**, 333–365, 1996. 428
7. Quinlan, J. R., *C4.5 - Programs for Machine Learning*, Morgan Kaufmann, Palo Alto, 1993. 432
8. *Readings in Machine Learning*, (Shavlik, J. W. and Dietterich, T. G., eds.) Morgan Kaufmann, Palo Alto, 1990. 423
9. 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. 425
10. Tsumoto, S., Automated Induction of Medical Expert System Rules from Clinical Databases based on Rough Set Theory. *Information Sciences* **112**, 67-84, 1998. 423, 432
11. Tsumoto, S. Extraction of Experts' Decision Rules from Clinical Databases using Rough Set Model *Intelligent Data Analysis*, 2(3), 1998. 429, 432

12. 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. 428
13. Ziarko, W., Variable Precision Rough Set Model. *Journal of Computer and System Sciences*. 46, 39-59, 1993.