Extraction of Experts' Decision Process from Clinical Databases Using Rough Set Model

Shusaku Tsumoto

Department of Information Medicine,
Medical Research Institute, Tokyo Medical and Dental University
1-5-45 Yushima, Bunkyo-ku Tokyo 113 Japan
E-mail: tsumoto.com@mri.tmd.ac.jp

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 is 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 on rule induction methods is that they cannot extract rules, which plausibly represent experts' decision processes[8].

Most rule induction methods induce probabilistic rules, the description length of which is too short, compared with the experts' rules. For example, rule induction methods, including AQ15[3] and PRIMEROSE[7], induce the following common rule for muscle contraction headache from databases on differential diagnosis of headache[8]:

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

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

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

where [Tenderness of B1=no] and [Tenderness of C1=no] are added. 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 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 subrules, rules discriminating between each group and rules classifying each class in the group are induced. Finally, those two parts are integrated into one rule for each decision attribute. The proposed method is evaluated on medical databases, the experimental results of which show that induced rules correctly represent experts' decision processes. The paper is organized as follows: in Section 2, we make a brief description about rough set theory and the definition of probabilistic rules based on this theory. Section 3 discusses interpretation of medical experts' rules. Then, Section 4 presents an induction algorithm for incremental learning. Section 5 gives experimental results. Section 6 discusses the problems of our work and related work, and finally, Section 7 concludes our paper.

2 Rough Set Theory and Probabilistic Rules

2.1 Rough Set Theory

Rough set theory clarifies set-theoretic characteristics of the classes over combinatorial patterns of the attributes, which are precisely discussed by Pawlak [4, 9]. This theory can be used to acquire some sets of attributes for classification and can also evaluate how precisely the attributes of database are able to classify data.

Let us illustrate the main concepts of rough sets which are needed for our formulation. Table 1 is a small example of database which collects the patients who complained of headache. First, let us consider how an attribute "loc" classify the headache patients' set of the table. The set whose value of the attribute "loc" is equal to "who" is {2,4,5,6}, which shows that the second, fourth, fifth and sixth case (In the following, the numbers in a set are used to represent each record number). This set means that we cannot classify {2,4,5,6} further solely by using the constraint $R = [loc = who]$. This set is defined as the indiscernible set over the relation R and described as follows: $[x]_R = \{2, 4, 5, 6\}$. In this set, {2,5} suffer from muscle contraction headache("m.c.h."), {4} from classical migraine("migraine"), and {6} from psycho("psycho"). Hence we need other additional attributes to discriminate between "m.c.h.", "migraine", and "psycho". Using this concept, we can evaluate the classification power of each attribute. For example, "nat=thr" is specific to the case of classic migraine ("migraine"). We can also extend this indiscernible relation to multivariate cases, such as $[x]_{[loc=who]\wedge[nau=0]} = \{2, 5\}$ and $[x]_{[loc=who]\vee[nat=no]} = \{1, 2, 4, 5, 6\}$, where $\wedge$ and $\vee$ denote "and" and "or" respectively. In the framework of rough set theory,

Table 1. An Example of Database

	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.
3	40-49	lat	thr	1	1	0	migra
4	40-49	who	thr	1	1	0	migra
5	40-49	who	rad	0	0	1	m.c.h.
6	50-59	who	per	0	1	1	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, 1: Yes, 0: No.

the set $\{2,5\}$ is called *strictly definable* by the former conjunction, and also called *roughly definable* by the latter disjunctive formula. Therefore, the classification of training samples D can be viewed as a search for the best set $[x]_R$ which is supported by the relation R. In this way, we can define the characteristics of classification in the set-theoretic framework. For example, accuracy and coverage, or true positive rate can be defined as:

$$\alpha_R(D) = \frac{|[x]_R \cap D|}{|[x]_R|}, \text{ and } \kappa_R(D) = \frac{|[x]_R \cap D|}{|D|},$$

where $|A|$ denotes the cardinality of a set A, $\alpha_R(D)$ denotes an accuracy of R as to classification of D, and $\kappa_R(D)$ denotes a coverage, or a true positive rate of R to D, respectively. For 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 $R \leftrightarrow D$.

For further information on rough set theory, readers could refer to [4, 9, 10].

2.2 Probabilistic Rules

The simplest probabilistic model is that which only uses classification rules which have high accuracy and high coverage.

This model is applicable when rules of high accuracy can be derived. Such rules can be defined as:

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

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

$$[M1 = 1] \rightarrow m.c.h.\ \alpha = 3/4 = 0.75,\ \kappa = 1.0,$$
$$[nau = 0] \rightarrow m.c.h.\ \alpha = 3/3 = 1.0,\ \kappa = 1.0,$$

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

It is notable that this rule is a kind of probabilistic proposition with two statistical measures, which is one kind of an extension of Ziarko's variable precision model(VPRS) [10]. [1]

3 Interpretation of Medical Experts' Rules

As shown in Section 1, rules acquired from medical experts are much longer than those induced from databases the decision attributes of which are given by the same experts. Those characteristics of medical experts' rules are fully examined not by comparing between those rules for the same class, but by comparing experts' rules with those for another class. For example, a classification rule for muscle contraction headache is given by:

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

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

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

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

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

[1] In VPRS model, the two kinds of precision of accuracy is given, and the probabilistic proposition with accuracy and two precision conserves the characteristics of the ordinary proposition. Thus, our model is to introduce the probabilistic proposition not only with accuracy, but also with coverage.

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 are 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.

4 Rule Induction

Rule induction 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.

4.1 An Algorithm for Rule Induction

An algorithm for rule induction is given as follows.

1. Calculate $\alpha_R(D)$ and $\kappa_R(D)$ for each elementary relation R and each class D.
2. Make a list of R $L(D)$ the coverage of which is equal to 1.0 ($L(D) = \{R|\kappa_R(D) = 1.0\}$) for each class D.
3. For each class D, make a list $L_2(D)$, each element $L(D_j)$ of which is a subset of $L(D)$.
4. Make a new decision attribute D' for each $L_2(D)$ and search for a partition P of all the classes D such that $L_2(D_i) \cap L_2(D_j) \neq \phi$.
5. Construct a new table $(T(P))$for P. Also construct a new table$(T(D'))$ for each decision attribute D'.
6. Induce classification rules R_p for each P in $T(P)$.
7. Induce classification rules R_d for each D in $T(D')$.
8. Integrate R_p and R_d into a rule $R(D)$.

Induction of Classification Rules For induction of classification rules, the algorithm introduced in PRIMEROSE[7] is applied, which is shown in Fig. 1.

Integration of Rules An algorithm for integration is given as follows.

1. For each D_i, repeat the following step.
2. Select one rule $R \rightarrow D_i$.
3. Search for a rule in D_i, $R' \rightarrow d_i$, the supporting set of which is a subset of that of $R \rightarrow D_i$.
4. Integrate these two rules into one: $R \wedge R' \rightarrow d_i$.

```
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 ( $L_i \neq \{\}$ ) do
     begin
      Select one pair $R = \wedge[a_i = v_j]$ from $L_i$;
      $L_i := L_i - \{R\}$;
      if   $(\alpha_R(D) \geq \delta_\alpha)$   and   $(\kappa_R(D) \geq \delta_\kappa)$
        then  do $S_{ir} := S_{ir} + \{R\}$; /* Include $R$ as Inclusive Rule */
      else $M := M + \{R\}$;
     end
    $L_{i+1}$ := (A list of the whole combination of the conjunction formulae in $M$);
   end
 end {Induction of Classification Rules };
```

Fig. 1. An Algorithm for Classification Rules

4.2 Example

Let us illustrate how the introduced algorithm works by using a small database in Table 1. For simplicity, the threshold δ_α is set to 1.0, which means that only deterministic rules should be induced.

After the first and second step, the following three $L(D_i)$ will be obtained: $L(m.c.h.) = \{[prod = 0], [M1 = 1]\}$, $L(migra) = \{[age = 40-49], [nat = who], [prod = 1], [nau = 1], [M1 = 0]\}$, and $L(psycho) = \{[age = 50-59], [loc = who], [nat = per], [prod = 0], [nau = 0], [M1 = 1]\}$.

Thus, since a relation $L(psycho) \subset L(m.c.h.)$ holds, 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.

$$\begin{array}{lll}
[M1 = 1] & \rightarrow D_1\ \alpha = 1.0,\ \kappa = 1.0, & \text{supported by } \{1,2,5,6\} \\
[prod = 0] & \rightarrow D_1\ \alpha = 1.0,\ \kappa = 1.0, & \text{supported by } \{1,2,5,6\} \\
[nau = 0] & \rightarrow D_1\ \alpha = 1.0,\ \kappa = 0.75, & \text{supported by } \{1,2,5\} \\
[nat = per] & \rightarrow D_1\ \alpha = 1.0,\ \kappa = 0.75, & \text{supported by } \{1,2,6\} \\
[loc = who] & \rightarrow D_1\ \alpha = 1.0,\ \kappa = 0.75, & \text{supported by } \{2,5,6\} \\
[age = 50-59] & \rightarrow D_1\ \alpha = 1.0,\ \kappa = 0.5, & \text{supported by } \{2,6\}
\end{array}$$

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.*.

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

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

$$\begin{array}{lll} [nau = 0] & \rightarrow m.c.h.\ \alpha = 1.0,\ \kappa = 1.0, & \text{supported by } \{1,2,5\} \\ [age = 40-49] & \rightarrow m.c.h.\ \alpha = 1.0,\ \kappa = 0.67, & \text{supported by } \{2,5\} \end{array}$$

In the eighth step, these two kinds of rules are integrated in the following way. For a rule $[M1 = 1] \rightarrow D_1$, $[nau = 0] \rightarrow m.c.h.$ 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:

$$\begin{array}{lll} [M1 = 1]\ \&\ [\text{nau=0}] & \rightarrow m.c.h.\ \alpha = 1.0,\ \kappa = 1.0, & \text{supported by } \{1,2,5\} \\ [M1 = 1]\ \&\ [\text{age=40-49}] & \rightarrow m.c.h.\ \alpha = 1.0,\ \kappa = 0.67, & \text{supported by } \{2,5\} \end{array}$$

5 Experimental Results

The above rule induction algorithm is implemented in PRIMEROSE4 (Probabilistic Rule Induction Method based on Rough Sets Ver 4.0), [2] and is applied to databases on headache and cerebrovascular diseases (CVD), whose precise information is given in Table 4.

This system is compared with PRIMEROSE [7], C4.5[5], CN2[1] and AQ15 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 is measured by the number of attribute-value pairs used in an induced rule and Jaccard's coefficient is adopted as a similarity measure, the definition of which is shown in the Appendix. Concerning the performance of rules, ten-fold cross-validation is applied to estimate classification accuracy.

[2] The program is implemented by using SWI-prolog [6] on Sparc Station 20.

Table 4. Information about Databases

Domain	Samples	Classes	Attributes
headache	1477	10	20
CVD	261	6	27

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

Table 5. Experimental Results

Method	Length	Similarity	Accuracy
	Headache		
PRIMEROSE4	8.6 ± 0.27	0.93 ± 0.08	93.3 ± 2.7%
Experts	9.1 ± 0.33	1.00 ± 0.00	98.0 ± 1.9%
PRIMEROSE	5.3 ± 0.35	0.54 ± 0.05	88.3 ± 3.6%
C4.5	4.9 ± 0.39	0.53 ± 0.10	85.8 ± 1.9%
CN2	4.8 ± 0.34	0.51 ± 0.08	87.0 ± 3.1%
AQ15	4.7 ± 0.35	0.51 ± 0.09	86.2 ± 2.9%
	CVD		
PRIMEROSE4	7.6 ± 0.37	0.89 ± 0.05	91.3 ± 3.2%
Experts	8.5 ± 0.43	1.00 ± 0.00	92.9 ± 2.8%
PRIMEROSE	4.3 ± 0.35	0.69 ± 0.05	84.3 ± 3.1%
C4.5	4.0 ± 0.49	0.65 ± 0.09	79.7 ± 2.9%
CN2	4.1 ± 0.44	0.64 ± 0.10	78.7 ± 3.4%
AQ15	4.2 ± 0.47	0.68 ± 0.08	78.9 ± 2.3%

6 Conclusion

In this paper, the characteristics of experts' rules are closely examined and a new approach to extract plausible rules is introduced. The proposed method is evaluated on medical databases, the experimental results of which show that induced rules correctly represent experts' decision processes.

References

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

2. Everitt. *Cluster Analysis*, 3rd Edition, 1996.
3. Michalski, R. S., Mozetic, I., Hong, J., and Lavrac, N. (1986). The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains. *Proceedings of the fifth National Conference on Artificial Intelligence*, 1041-1045, AAAI Press, Palo Alto, CA.
4. Pawlak, Z. (1991). *Rough Sets.* Kluwer Academic Publishers, Dordrecht.
5. Quinlan, J.R. (1993). *C4.5 - Programs for Machine Learning*, Morgan Kaufmann, CA.
6. SWI-Prolog Version 2.0.9 Manual, (1995). University of Amsterdam.
7. Tsumoto, S. and Tanaka, H. PRIMEROSE: Probabilistic Rule Induction Method based on Rough Sets and Resampling Methods. *Computational Intelligence*, **11**, 389-405, 1995.
8. Tsumoto, S. Empirical Induction of Medical Expert System Rules based on Rough Set Model. PhD dissertation, 1997 (in Japanese).
9. Ziarko, W. (1991). The Discovery, Analysis, and Representation of Data Dependencies in Databases. in: Shapiro, G. P. and Frawley, W. J. (eds), *Knowledge Discovery in Databases*, AAAI press, Palo Alto, CA, pp.195-209.
10. Ziarko, W. (1993). Variable Precision Rough Set Model. *Journal of Computer and System Sciences*, **46**, 39-59.

A Appendix

PRIMEROSE4 calculates the following similarity measure from all the inputs. Although there are many kinds of similarities[2], a family of similarity measures based on a contingency table is adopted. Let us consider a contingency table for a rule of a certain disease (Table 6). The first and second column denote the positive and negative information of an experts' rule. The first and second row denote the positive and negative information of an induced rule. Then, for example, a denotes the number of attributes in an induced rule which matches an experts' rule. 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 7. In PRIMEROSE4, users can choose a similarity measure from these four. As a default, Jaccard's coefficient, is used for defining similaritites,

Table 6. Contigency Table for Similarity

		Rule		
		1	0	Total
Sample	1	a	b	a+b
	0	c	d	c+d

Table 7. Definition of Similarity Measures

(1) Matching Number	a
(2) Jaccard's coefficient	$a/(a+b+c)$
(3) χ^2-statistics	$N(ad-bc)^2/M$
(4) point correlation coefficient	$(ad-bc)/\sqrt{M}$

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

because it satisfies not only the low computational complexity, but also a good performance.