Automated Knowledge Acquisition from Clinical Databases based on Rough Sets and Attribute-Oriented Generalization

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@computer.org

Rule induction methods have been proposed in order to acquire knowledge automatically from databases. However, conventional approaches do not focus on the implementation of induced results into an expert system. In this paper, the author focuses not only on rule induction but also on its evaluation and presents a systematic approach from the former to the latter as follows. First, a rule induction system based on rough sets and attribute-oriented generalization is introduced and was applied to a database of congenital malformation to extract diagnostic rules. Then, by the use of the induced knowledge, an expert system which makes a differential diagnosis on congenital disorders is developed. Finally, this expert system was evaluated in an outpatient clinic, the results of which show that the system performs as well as a medical expert.

INTRODUCTION

Rule induction methods have been proposed in order to acquire knowledge automatically from databases[1,2]. They are expected to solve the bottleneck problem of knowledge acquisition[3] and good performances of induced rules are reported since the beginning of 1980s[1]. However, no paper focuses on the implementation of induced results into an expert system.

In this paper, the author focuses not only on rule induction but also its evaluation, and presents a systematic approach from introduction of automatic knowledge acquisition methods to evaluation of the induced rules in clinical environment in the following four steps. Firstly, the author focuses on two kinds of medical reasoning, called positive and negative reasoning and defines two corresponding rules, positive and negative rules, in the framework of rough sets[4]. Also, in order to extract more generalized rules, the author introduces attribute-oriented generalization[5], which converts attributes given in a database into generalized ones by using domain knowledge. Secondly, according to the formal definition of positive, negative and generalized rules, a rule induction system, called PRIMEROSE-REX2 (Probabilistic Rule Induction Method based on Rough Sets for Rules of Expert System ver 2.0), is developed. This system was applied to a clinical database of congenital malformation[6,7] in order to acquire diagnostic knowledge. The induced results show that this method not only extracts experts' knowledge correctly, but also discovers that symptoms observed in six positions (eyes, noses, ears, lips, fingers and feet) play important roles in a diagnosis. Thirdly, according to the induced rules and attribute information, the author develops an expert system, called RCOM (Rule based system for Congenital Malformation), which outputs a diagnostic conclusion (a syndrome) from the observations input by users. The system provides a graphical input interface for symptoms in six positions which consists of a list of photographs and verbal input interfaces for other symptoms. Each photograph describes typical patterns of symptoms in eyes, noses, ears, lips, fingers and feet so that users will not miss inputs necessary for diagnostic procedures. After all the inputs are completed, this system applies all the inputs to diagnostic rules and outputs a congenital disorder as a diagnostic conclusion. Finally, this expert system was evaluated in clinical practice, the results of which show that the system gains as good performance as a medical expert.

FOCUSING MECHANISM

One of the characteristics in medical reasoning is a focusing mechanism, which is used to select the final diagnosis from many candidates[2]. For example, in a differential diagnosis of headache, more than 60 diseases should be checked. This style of reasoning consists of the following two kinds of reasoning processes: negative reasoning and positive reasoning, which are applied in the following way. First, negative reasoning excludes a disease from candidates when a patient does not have a symptom which is necessary to diagnose that disease. Then, positive reasoning suspects a disease in the output of the exclusive process when a patient has symptoms specific to a disease. These two steps are modeled as usage of two kinds of rules, negative rules (exclusive rules) and positive rules, the former and the latter of which corresponds to negative reasoning and positive reasoning, respectively. In the next section, these two rules are represented as special kinds of probabilistic rules.

DEFINITION OF RULES

In the following sections, the author uses the following notations of rough set theory[4], which is illustrated by a small database shown in Table 1. First, a combination of attribute-value pairs, corresponding to a complex in AQ terminology[1] is denoted by a formula R. For example, *[telorism=hyper]* $\wedge$ *[iris-defects=yes]* will be one formula, denoted by $R=$*[telorism=hyper]*$\wedge$ *[iris-defects=yes]*. Secondly, a set of samples which satisfy R is

1091—8280/98/$5.00 © 1998 AMIA, Inc.

denoted by $[x]_R$, corresponding to a star in AQ terminology. For example, the set, $[x]_{[telorism=hyper]}$, each member of which satisfies *[telorism=hyper]*, is equal to {2,4,5,6}, which shows that the second, fourth, fifth and sixth case satisfy the above relation (In the following, the numbers in a set are used to represent each record number). This notation can be also extended to a conjunctive or disjunctive formula, such as $[x]_{[telorism=hyper]\wedge[iris\text{-}defects=yes]}=$ {2,5} and$[x]_{[telorism=hyper]\vee[cornea=no]}=$ {1,2,4,5,6}, where ∧and ∨ denote "and" and "or" respectively. Finally, *U*, which stands for "Universe", denotes all training samples.

Table 1: A Small Database on Congenital Disorders

U	A1	A2	A3	A4	A5	A6	Class
1	No	Normal	Megalo	Yes	Yes	Long	Aarskog
2	Yes	Hyper	Megalo	Yes	Yes	Long	Aarskog
3	Yes	Hypo	Normal	No	No	Normal	Down
4	Yes	Hyper	Normal	Yes	No	Normal	Down
5	Yes	Hyper	Large	Yes	Yes	Long	Aarskog
6	No	Hyper	Megalo	Yes	No	Long	Cat-cry

Definitions: A1: round face, A2: telorism, A3:cornea, A4: antimongoloid slanting of palpebral fissures, A5: iris-defects, A6: eyelashes, Aarskog: Aarskog syndrome, Down: Down syndrome, Cat-cry: Cat-cry syndrome.

Accuracy and Coverage

By the use of these notations, classification accuracy and coverage (true positive rate) are defined as:

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

where $[x]_R$ and $|A|$ denote a set which satisfy a relation *R*, the cardinality of a set *A*, respectively. In the above example, when *R* and *D* are set to *[iris-defects=no]* and *[class=Down]*, the accuracy and coverage of *R* as to *D* are calculated as follows: $\alpha_R(D)=2/3=0.67$ 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→D and that $\kappa_R(D)$ measures the degree of its necessity. For example, if the accuracy is equal to 1.0, then R→D is true, and if the coverage is equal to 1.0, then D→R is true. Thus, if both measures are equal to 1.0, then R ↔ D will hold.

Probabilistic Rules

By the use of accuracy and coverage, a probabilistic rule is defined as:

$$R \to d, R = \wedge_i[a_i = v_j], \alpha_R(D) \geq \delta_\alpha, \kappa_R(D) \geq \delta_\kappa,$$

which is an extension of rules in Ziarko's variable precision rough set model[8].

This type of rule is mainly used to represent rules which are induced after the application of attribute-oriented generalization. It is notable that both positive rules and negative rules are defined as special cases of this rule, as shown in the next subsections.

Positive Rules

A positive rule is defined as a rule supported by only positive examples, the classification accuracy of which is equal to 1.0. It is notable that the set that supports this rule corresponds to a subset of the lower approximation of a target concept, which is introduced in rough sets[4]. A positive rule is represented as:

$$R \to d, R = \wedge_i[a_i = v_j], \alpha_R(D) = 1.0.$$

In the above example, one positive rule of ``Aarskog" syndrome is:

[iris-defects=yes] → Aarskog, $\alpha=3/3=1.0$.

Positive rules are often called deterministic rules. However, in this paper, a term, positive (deterministic) rules is used, because deterministic rules which is supported only by negative examples, called negative rules, are also introduced as in the next subsection.

Negative Rules

Before defining a negative rule, let us first introduce an exclusive rule, the contrapositive of a negative rule[2]. An exclusive rule is defined as a rule supported by all the positive examples, the coverage of which is equal to 1.0. (Exclusive rules represent the necessity condition of a decision.) It is notable that the set supporting a exclusive rule corresponds to the upper approximation of a target concept, introduced in rough sets.

An exclusive rule is represented as:

$$R \to d, R = \vee[a_i = v_j], \kappa_R(D) = 1.0.$$

In the above example, an exclusive rule of ``Aarskog" is:

[eyelashes=long] ∨[iris-defects=yes] → m.c.h. , $\kappa=1.0$.

From the viewpoint of propositional logic, an exclusive rule should be represented as:

$$d \to \vee_i[a_i = v_j],$$

because the condition of an exclusive rule corresponds to the necessity condition of a conclusion *d*. Thus, it is easy to see that a negative rule is defined as the contrapositive of an exclusive rule:

$$\wedge_i \neg[a_i = v_j] \to \neg d$$

which means that if a case does not satisfy any attribute value pairs in the condition of a negative rule, then a decision *d* should be excluded from diagnostic candidates. For example, the negative rule of Aarskog is:

$\wedge\neg$[eyelashes=long]$\wedge\neg$[iris-defects=yes]$\to\neg$Aarskog, In summary, a negative rule is defined as:

$$\wedge_i \neg[a_i = v_j] \rightarrow \neg d \qquad s.t. \qquad \forall[a_i = v_j], \kappa_{[a_i = v_j]}(D) = 1.0,$$

where D denotes a set of samples which belong to a class d. Negative rules should be also included in a category of deterministic rules, since their coverage, a measure of negative concepts is equal to 1.0.

Attribute-Oriented Generalization

Rule induction methods regard a database as a decision table[4] and induce rules, which can be viewed as reduced decision tables. However, those rules extracted from tables do not include information about attributes and they are too simple. In a practical situation, domain knowledge about attributes is very important to gain the comprehensability of induced knowledge, which is one of the reasons why databases are implemented as relational-databases[5].

Thus, reinterpretation of induced rules by using information about attributes is needed to acquire comprehensive rules. For example, terolism, cornea, antimongoloid slanting of palpebral fissures, iris defects and long eyelashes are symptoms around eyes. Thus, those symptoms can be gathered into a category "eye symptoms" when the location of symptoms should be focused on. This process, grouping of attributes, is called attribute-oriented generalization[5].

Attribute-oriented generalization can be viewed as transformation of variables in the context of rule induction. For example, an attribute "iris defects" should be transformed into an attribute "eye symptoms". (It is notable that the transformation of attributes in rules corresponds to that of a database because a set of rules is equivalent to a reduced decision table.) So, one positive rule of ``Aarskog",

$$[\text{iris-defects=yes}] \rightarrow \text{Aarskog}, \quad \alpha=3/3=1.0$$

is rewritten as:

$$[\text{eye-symptoms=yes}] \rightarrow \text{Aarskog}.$$

Since five attributes (telorism, cornea, slanting, iris-defects and eyelashes) are generalized into *eye-symptoms*, the candidates for accuracy and coverage will be (5/6, 2/3), (3/4, 3/3), (3/4, 3/3), (3/3, 3/3), and (3/4, 3/3), respectively. In these values, minimum one should be selected: accuracy is equal to 3/4 and coverage is equal to 2/3. Thus, the rewritten rule becomes the following probabilistic rule:

$$[\text{eye-symptoms=yes}] \rightarrow \text{Aarskog}, \alpha=3/4=0.75, \kappa=2/3=0.67.$$

This process gives us information about the location to which medical experts pay attention in order to describe a syndrome. In the case of Aarskog syndrome, all the positive rules given below show that eye symptoms are very important for its diagnosis.

ALGORITHMS

Induction of Negative Rules

The contrapositive of a negative rule, an exclusive rule is induced as an exclusive rule by the modification of the algorithm introduced in PRIMEROSE-REX[2], as shown in Fig. 1. Negative rules are derived as the contrapositive of induced exclusive rules.

```
procedure Exclusive and Negative Rules;
  var
     L, $L_{ir}$: List ; /* A list of elementary attribute-value pairs */
  begin
     L:= P0;   /* P0: A list of elementary attribute-value pairs
                    given in a database */
     while (L ≠{}) do
       begin
          Select one pair $[a_i=v_j]$ from L;
          if ($[x]_{[ai=vj]} \cap D \neq \emptyset$) then  do
                /* D: positive examples of a target class d */
             begin
                $L_{ir}$:= $L_{ir}$+$[a_i=v_j]$ ;
                     /* Candidates for Positive Rules */
                if ($\kappa_{[ai=vj]}(D)$=1.0)
                     then   $R_{er}$:= $R_{er} \vee [a_i=v_j]$;
                /* Include $[a_i=v_j]$ in the Exclusive Rule */
             end
           L:=L-$[a_i=v_j]$ ;
       end
    Construct Negative Rules: take the contrapositive of $R_{er}$;
  end {Exclusive and Negative Rules};
```

Figure 1: Induction of Exclusive and Negative Rules

Induction of Positive Rules

Positive rules can be viewed as a specific type of inclusive rules introduced in a rule induction system, PRIMEROSE-REX[2], the accuracy and coverage of which is equal to 1.0 and 0.0, respectively. Both rules can be induced by using the algorithm in PRIMEROSE-REX, shown in Fig. 2. For induction of positive rules, we only have to change the thresholds of accuracy and coverage into 1.0 and 0.0, respectively.

Rule Induction for Generalized Attributes

After induction of positive and negative rules, attributes are transformed into generalized ones, according to the list given by users. Each element of the list is represented as a tuple, *(a_i,A,C)*, where a_i, A and C denote an attribute in a given data set, a generalized attribute for a_i, and the upper level concept of A, respectively. For example, *(iris-defects, eye-symptoms, Location)* means the iris-defect is generalized into eye-symptoms with respect to the location of symptoms. Then, attributes in both rules are transformed by using generalized attributes *(A,C)* and statistics of generalized attributes are obtained from induced rules(Fig. 3).

PRIMEROSE-REX2

The author develops a rule induction system, called PRIMEROSE-REX2 (Probabilistic Rule Induction Method based on ROugh SEt for Rules of Expert System ver 2.0) by using the introduced algorithms. This system automatically acquires knowledge in the following way. First, PRIMEROSE-REX2 induces negative and positive rules from a given database. Then, it applies attribute-oriented generalization to the induce rules and changes attributes to generalized ones.(This process is equivalent to transformation of variables in a database.) Finally, the system calculates the statistics of rules obtained.

```
procedure Positive Rules;
    var
        i: integer;    M, $L_i$: List;
    begin
        $L_1$:= $L_{ir}$;   /* $L_{ir}$: A list of candidates generated by
                                induction of exclusive rules */
        i:=1; M:={};
        for i:=1 to n do
            /* n: Total number of attributes given in a database */
          begin
            while ($L_i \neq${}) do
              begin
                Select one pair R=$\wedge$ [ai=vj] from $L_i$ ;
                $L_i$ := $L_i$ - {R} ;
                if ($\alpha_R(D) > \delta_\alpha$) then do
                                      $S_{ir}$ := $S_{ir}$ + {R};
              /* Include R in a list of the Positive Rules */
                else M := M + {R};
              end
            $L_{i+1}$:= (A list of the whole combination of
                          the conjunction formulae in M);
          end
    end {Positive Rules};
```

Figure 2: Induction of Positive Rules

```
procedure Generalized Rules;
    var
        i: integer;    M, L_i: List;
        $L_{gen}$: List; /* List for Attribute-Oriented Generalization */
    begin
      while ( $L_{gen} \neq$ {} )    do
            begin
                Select one pair ($a_i$,A,C) from $L_i$;
                    /* A: generalized attribute for $a_i$ */
                    /* C: General Concepts of    A */
                Change $a_i$ in Induced Rules into (A,C);
                $L_i$:= $L_i$ - {($a_i$,A,C)};
            end
          Calculate Statistics for Generalized Rules for each C;
    end {Generalized Rules};
```

Figure 3.   Induction of Generalized Rules

EXPERIMENTAL RESULTS

PRIMEROSE-REX2 was applied to a clinical database of congenital malformation, which consists of 336 samples, 268 attributes and 12 diseases. Furthermore, attributes are classified into 12 general attributes (head, hair, face, eyes, nose, ears, lips, body, arms, fingers, legs and feet).

Statistics of Induced Rules

As statistics, the number and length of induced rules are used and compared with those acquired directly from the expert in Tokyo Medical and Dental University. Concerning the number of rules for each disease, the expert gives one positive and negative rule for each congenital disorder.

The results obtained are summarized into Table 2, which shows the averaged number and length of the positive and negative rules induced by PRIMEROSE-REX2. Concerning postive rules, the averaged number and length of rules obtained is 2.83 and 6.63, respectively, while the averaged length of expert's rules is 6.66, which suggests that this system induce rules similar to the expert of congenital disorders. On the other hand, the averaged length of negative rules obtained is 10.67, whereas the averaged length of expert's rules is 7.41. Interestingly, the negative rules acquired from the database are more redundant than expert's rules.

Table2: Number and Length of Positive and Negative Rules

Rules	Samples	Number of Rules	Length of Rules	Expert
Positive Rules	336	2.83±0.75	6.63±1.21	6.66±0.94
Negative Rules	336	1.0±0.0	10.67±3.45	7.41±2.11

Statistics of Generalized Attributes

As a statistic, the total number of generalized attributes in positive and negative rules is selected to see which generalized attributes are important for a diagnostic procedure. Table 3 shows the total number of generalized attributes used in positive and negative rules. As a result, symptoms in eyes, noses, ears, lips, fingers and feet are frequently used to describe those rules, which suggest that these six locations should be indispensable to make a differential diagnosis.

Table 3. Number of Generalized Attributes used in Both Rules

Location	Positive Rules	Negative Rules
Head	0.50±0.12	0.25±0.09
Hair	0.33±0.08	0.33±0.21
Face	0.58±0.11	0.42±0.05
Eyes	**1.75±0.33**	**3.58±0.86**
Nose	**1.08±0.75**	**1.67±0.38**
Ears	**0.83±0.12**	**0.83±0.41**
Lips	**0.75±0.04**	**0.75±0.39**
Body	0.33±0.04	0.33±0.08
Arms	0.33±0.02	0.50±0.13
Fingers	**1.08±0.43**	**0.92±0.21**
Legs	0.33±0.09	0.17±0.07
Feet	**0.75±0.12**	**0.83±0.20**

Expert System: RCOM

An expert system, called RCOM (Rule-based system for COngenital Malformation) is developed by using the acquired knowledge. The most important induced results are that symptoms in eyes, noses, ears, lips, fingers and feet are indispensable to make a differential diagnosis. Thus, lists of photographs, each of which describes typical patterns of symptoms, are used to construct input-interfaces for six locations (eyes, noses, ears, lips, fingers and feet) so that users do not miss any inputs (symptoms) important for diagnosis. Users select one of those photographs in each location, which is similar to a patient in his/her outpatient clinic. After users select the photographs in six locations, RCOM retrieves the symptoms from each photograph as inputs, then applies them to positive and negative rules induced by PRIMEROSE-REX2 and finally outputs diagnostic conclusions.

Evaluation of RCOM

RCOM was evaluated in clinical practice with respect to its classification accuracy by using 93 patients who came to the outpatient clinic after the development of this system. Experimental results about classification accuracy are shown in Table 4. The first and second row show the performance of rules obtained by using PRIMROSE-REX2: the results in the first row are derived by using both positive and negative rules and those in the second row are derived by only positive rules. The third and fourth row show the results derived by using both positive and negative rules and those by positive rules acquired directly from a medical expert in Tokyo Medical and Dental University. These results show that the combination of positive and negative rules outperforms positive rules and gains almost the same performance as that expert .

Table 4: Experimental Results

Method	Accuracy
PRIMEROSE-REX2 (Positive and Negative)	91.4 % (85/93)
PRIMEROSE-REX (Positive)	78.5% (73/93)
Experts (Positive and Negative)	93.5% (87/93)
Expert (Positive)	82.8% (77/93)

DISCUSSION

In this paper, a rule induction system based on rough sets and attribute-oriented generalization is introduced and was applied to a clinical database of congenital disorders. The experimental results above show that PRIMEROSE-REX2 automatically acquired diagnostic rules whose performance is as good as a medical expert. Interestingly, as shown in Table 4, the difference between RCOM and a medical expert in misclassification is only two cases: further analysis shows that these two cases are complicated and have at least two different diagnostic candidates, which means that these cases are very difficult even for a domain expert to diagnose. Thus, this system can be said to achieve almost the same performance as a medical expert of congenital disorders.

However, this discussion may be true in a differential diagnosis from 12 diseases, which are frequently observed in an outpatient clinic: actually, more than 1000 syndromes are reported in the domain of congenital malformation[6,7], some of which are rarely observed. The introduced knowledge acquisition system cannot support these syndromes unobserved in a given database. Thus, rules induced from databases should be reviewed by domain experts to check whether acquired knowledge covers a domain enough or not. If not, it is necessary to acquire knowledge from the experts about unobserved diseases or to ask them to check the quality of given databases. It will be our future work to develop a more sophisticated knowledge acquisition process in which rule induction methods help domain experts to extract useful information from databases.

CONCLUSIONS

In this paper, the author presents a systematic approach from automated knowledge acquisition method to evaluation of an medical expert system based on induced rules. First, positive, negative and generalized rules are defined by using rough sets and attribute-oriented generalization and a rule induction method is introduced. Secondly, the system was applied to a database of congenital malformation. Thirdly, by the use of the induced knowledge, an expert system which makes a differential diagnosis on congenital malformation is developed. Finally, this expert system was evaluated in clinical practice, the results of which show that the system performs as well as a medical expert.

Acknowledgements

The author would like to thank Kimie Ohyama for her help on construction of databases and her insightful comments.

References

1. Shavlik, JW and Dietterich, TG. (eds.) *Readings in Machine Learning*, CA: Morgan Kaufmann, 1990.

2. Tsumoto, S. and Tanaka, H. Automated Discovery of Medical Expert System Rules from Clinical Databases based on Rough Sets. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining 96.* CA: AAAI Press, 1996; 63-69.

3. Buchnan, BG and Shortliffe, EH.(eds.) *Rule-Based Expert Systems*. CA: Addison-Wesley, 1984.

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

5. Cai, YD, Cercone, N. and Han, J. Attribute-oriented induction in relational databases. In:
Shapiro, GP. and Frawley, WJ. (eds), *Knowledge Discovery in Databases*. CA: AAAI press,
1991; 213-228.

6. Behrman, R. (ed.) *Principles of Pediatrics*, 14th edition. Philadelphia: W.B. Saunders, 1992.

7. Goodman, RM, Golin, RJ. *Atlas of the face in genetic disorders.* 2nd edition, Saint Louis: C.V. Mosby, 1977.

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