Medical Reasoning and Rough Sets

Shusaku Tsumoto

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

Abstract. Pawlak showed that knowledge can be captured by data partition and proposed a rough set method where comparison between data partition gives knowledge about classification. Interestingly, thes approximations correspond to the focusing mechanism of differential medical diagnosis; upper approximation as selection of candidates and lower approximation as concluding a final diagnosis. This paper focuses on severl models of medical reasoning shows that core ideas of rough set theory can be observed in these diagnostic models.

1 Introduction

Medical reasoning always includes uncertainty[1], which is caused by the limitations of medical knowledge, available data and our recognition, compared with the complexities of human body. Thus, medical databases also have a certain degree of uncertainty: rules extracted from databases are also incomplete, which suggests that rule induction method should deal with uncertain rules.

According to this motivation, rule induction based on rough set theory have been applied to medical databases empirically[2,3], the results of which shows that rough-set-based methods are very useful to extract medical diagnostic rules.

This paper presents how medical diagnostic rules are modeled by the concepts of rough sets[4] in a more theoretical way. The key ideas are variable precision rough set model, which corresponds to a ordinal positive reasoning, and an upper approximation of a target concept, which corresponds to a focusing mechanism in medical reasoning. Acquired models show that the characteristics of medical reasoning reflect the concepts on approximation of rough sets, which explains why rough sets work well in medical domains. The paper is organized as follows: in Section 2, two important measures, accuracy and coverage are defined and a probabilistic rule is defined. Section 3 to 5 presents description of three types of medical reasoning: simple differential diagnosis, focusing mechanism and $m-$of$-n$ criteria, respectively. Section 6 concludes our paper.

2 Definition of Rules

2.1 Rough Sets

In the following sections, we use the following notations introduced by Grzymala-Busse and Skowron[5], which are based on rough set theory[4]. These notations

M. Kryszkiewicz et al. (Eds.): RSEISP 2007, LNAI 4585, pp. 90–100, 2007.

Table 1. An Example of Dataset

No.	age	location	nature	prodrome	nausea	M1	class
1	50-59	occular	persistent	no	no	yes	m.c.h.
2	40-49	whole	persistent	no	no	yes	m.c.h.
3	40-49	lateral	throbbing	no	yes	no	migra
4	40-49	whole	throbbing	yes	yes	no	migra
5	40-49	whole	radiating	no	no	yes	m.c.h.
6	50-59	whole	persistent	no	yes	yes	psycho

DEFINITIONS. M1: tenderness of M1, m.c.h.: muscle contraction headache, migra: migraine, psycho: psychological pain.

are illustrated by a small dataset shown in Table 1, which includes symptoms exhibited by six 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., $a : U \to V_a$ 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 f_A is equal to $g_{location,nausea} = \{2, 5\}$.

2.2 Classification Accuracy and Coverage

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

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 $R \to d$ is defined as:

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

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.

Figure 1 depicts the Venn diagram of relations between accuracy and coverage. Accuracy views the overlapped region $|R_A \cap D|$ from the meaning of a relation R. On the other hand, coverage views the overlapped region from the meaning of a concept D.

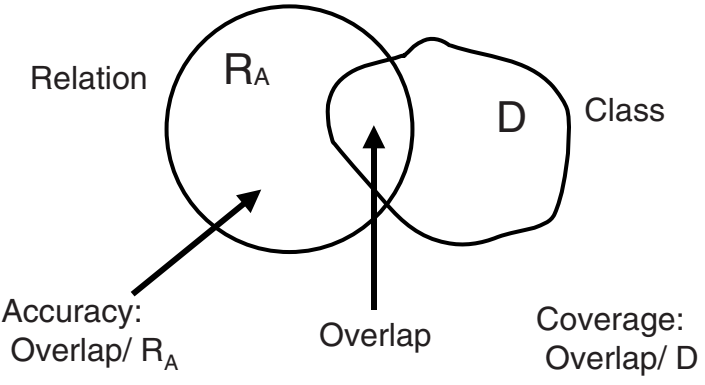

Fig. 1. Venn Diagram of Accuracy and Coverage

In the above example, when R and D are set to $[nau = yes]$ 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 $R \leftrightarrow D$. Other characteristics of accuracy and coverage are shown in the appendix.

2.3 Probabilistic Rules

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

$$R \stackrel{\alpha,\kappa}{\rightarrow} d \quad s.t. \ R = \wedge_j [a_j = v_k], \alpha_R(D) \geq \delta_\alpha \ \text{ and } \kappa_R(D) \geq \delta_\kappa,$$

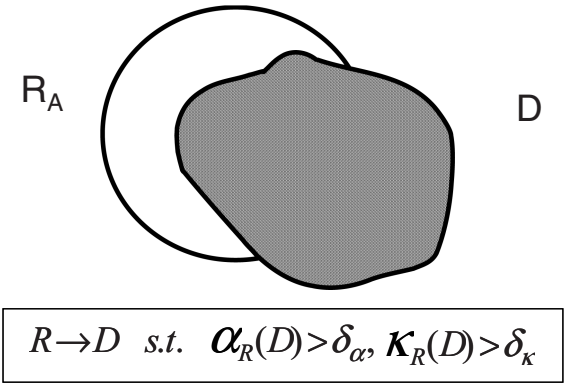

Fig. 2. Venn Diagram for Probabilistic Rules

If the thresholds for accuracy and coverage are set to high values, the meaning of the conditional part of probabilistic rules corresponds the highly overlapped region. Figure 2 depicts the Venn diagram of probabilistic rules with highly overlapped region. This rule is a kind of probabilistic proposition with two statistical measures, which is an extension of Ziarko's variable precision model(VPRS) [6].[1]

3 Simplest Diagnostic Rules

3.1 Representation of Diagnostic Rules

The simplest probabilistic model is that which only uses classification rules which have high accuracy and high coverage. Such rules can be defined as:

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

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 (both δ_α and δ_κ are set to 0.75):

$$\begin{array}{l} [prod = 0] \rightarrow m.c.h.\ \alpha = 3/4 = 0.75,\ \kappa = 1.0, \\ [nau = 0] \rightarrow m.c.h.\ \alpha = 3/3 = 1.0,\ \kappa = 1.0, \\ [M1 = 1] \rightarrow m.c.h.\ \alpha = 3/4 = 0.75,\ \kappa = 1.0, \end{array}$$

3.2 An Rule Induction Algorithm

An rule induction algorithm is defined as Figure 1, which is discussed precisely in [8]. It is notable that rule induction of other type rules is derived by simple modification of this algorithm.

[1] This probabilistic rule is also a kind of *Rough Modus Ponens*[7].

```
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
            Sort L_i with respect to the value of coverage;
            Select one pair R = ∧[a_i = v_j] from L_i,
            which have the largest value on coverage;
            L_i := L_i − {R};
            if (κ_R(D) ≥ δ_κ)
              then do
                      if (α_R(D) ≥ δ_α)
                        then  do S_ir := S_ir + {R}; /* Include R as Classification Rule */
                      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. 3. An Algorithm for Classification Rules

4 Focusing Mechanism

One of the characteristics in medical reasoning is a focusing mechanism, which is used to select the final diagnosis from many candidates[9,10]. For example, in differential diagnosis of headache, more than 60 diseases will be checked by present history, physical examinations and laboratory examinations. In diagnostic procedures, a candidate is excluded if a symptom necessary to diagnose is not observed.

This style of reasoning consists of the following two kinds of reasoning processes: exclusive reasoning and inclusive reasoning. Relations of this diagnostic model with another diagnostic model are discussed in [2]. The diagnostic procedure will proceed as follows (Figure 4): first, exclusive reasoning excludes a disease from candidates when a patient does not have a symptom which is necessary to diagnose that disease. Secondly, inclusive reasoning suspects a disease in the output of the exclusive process when a patient has symptoms specific to a disease. These two steps are modelled as usage of two kinds of rules, negative rules (or exclusive rules) and positive rules, the former of which corresponds to exclusive reasoning and the latter of which corresponds to inclusive reasoning. In the next two subsections, these two rules are represented as special kinds of probabilistic rules.

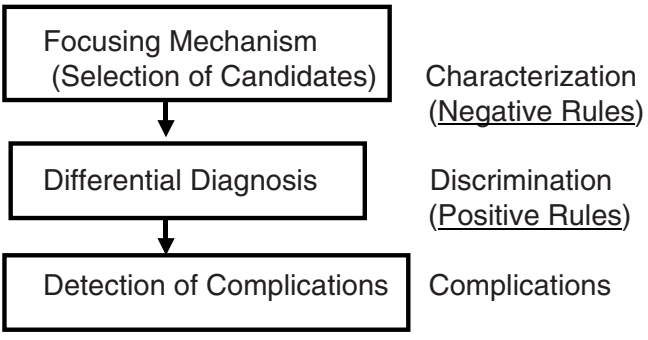

Fig. 4. Illustration of Focusing Mechanism

4.1 Positive Rules

A positive rule can be defined as a rule supported by only positive examples, which means that the classification accuracy of a rule is equal to 1.0. Thus, a positive rule is represented as:

$$R \rightarrow d \quad s.t. \qquad R = \wedge_j [a_j = v_k], \quad \alpha_R(D) = 1.0$$

In the above example, one positive rule of "m.c.h." is:

$$[nau = 0] \rightarrow m.c.h. \quad \alpha = 3/3 = 1.0.$$

This positive rule is often called deterministic rules. However, in this paper, we use a term, positive (deterministic) rules, because deterministic rules which is supported only by negative examples, called negative rules, is introduced as in the next subsection.

4.2 Negative Rules

Before defining a negative rule, let us first introduce an exclusive rule, the contrapositive of a negative rule[9]. An exclusive rule can be defined as a rule supported by all the positive examples, which means that the coverage of a rule is equal to 1.0.[2] Thus, an exclusive rule is represented as:

$$R \rightarrow d \quad s.t. \qquad R = \wedge_j [a_j = v_k], \quad \kappa_R(D) = 1.0.$$

In the above example, exclusive rule of "m.c.h." is:

$$[prod = 0] \wedge [nau = 0] \wedge [M1 = 1] \rightarrow m.c.h. \quad \kappa = 1.0,$$

It is notable that exclusive rule corresponds to an upper approximation of a target concept. For example, the set which supports the exclusive rule above is an upper approximation of m.c.h.

[2] Exclusive rules represent the necessity condition of a decision.

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

$$d \rightarrow \wedge_j [a_j = v_k],$$

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

$$\vee_j \neg [a_j = v_k] \rightarrow \neg d,$$

which means that if a case does not satisfy any attribute value pairs in the condition of a negative rules, then we can exclude a decision d from candidates. For example, the negative rule of m.c.h. is:

$$\neg [prod = 0] \vee \neg [nau = 0] \vee \neg [M1 = 1] \rightarrow \neg m.c.h.$$

In summary, a negative rule is defined as:

$$\wedge_j \vee [a_j = v_k] \rightarrow \neg d \quad s.t. \quad \forall [a_j = v_k] \ \kappa_{[a_j = v_k]}(D) = 1.0,$$

where D denotes a set of samples which belong to a class d. It can be also called a deterministic rule, since a measure of negative concept, coverage is equal to 1.0.

In summary, positive and negative rules corresponds to positive and negative regions defined in rough sets. Figure 5 shows the Venn diagram of those rules.

4.3 Rule Induction Algorithm

An algorithm for induction of positive and negative rules is derived by simple modification of the algorithm in Figure 1: if the thresholds of accuracy and coverage is set to 0.0 and 1.0, respectively, the algorithm for negative rules will

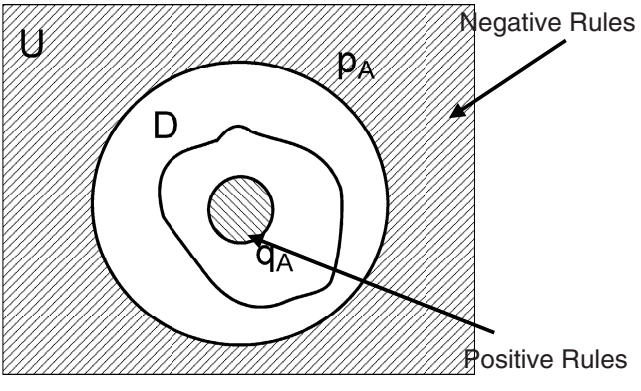

Fig. 5. Positive and Negative Rules as Overview

be obtained. On the other hand, if the thresholds of accuracy and coverage is set to 1.0 and 0.0, respectively, the algorithm for negative rules will be obtained.

It is notable that positive and negative rules can be extended to probabilistic versions, which is discussed precisely in [9].

5 Criteria Tables

5.1 Representation of Rules

Another characteristic reasoning in medicine is $m-$of$-n$ concepts, or criteria table, which is discussed in [11]. Criteria table for a disease d is described by n attributes, which are enough to make its diagnosis. If at least m attributes are observed in a patient, d should be suspected.

Langley discusses that this $m-$of$-n$ description can be rewritten as a simple linear combination of attribute-value pairs. Thus, he implements an induction of this description as an induction of threshold concepts.

However, a $m-$of$-n$ rule in medicine is not equivalent to a linear combination rule, which is a special kind of statistical discriminant functions[12]. Rather, this type of rule is based on relations between sets as follows.

1. If total n attributes are observed, a disease d is suspected with the highest accuracy. (The coverage is equal to 1.0).

2. If m attributes are satisfied, a disease d should be suspected with high accuracy. (The coverage is equal to 1.0).

3. If less than m attributes are satisfied, the probability of d is low. However, the coverage is equal to 1.0. Thus, $m-$of$-n$ concept is described as combination of exclusive rules (below, we call them *unit rules*) with the constraint that their accuracies are high:

$$R \to d \; s.t. \; R = \wedge_{j=1}^{i}[a_j = v_k](m \leq i \leq n)$$
$$\alpha_R(D) \geq \delta_\alpha, \kappa_{[a_j=v_k]}(D) = 1.0,$$

which also satisfies that: if R is represented as $\wedge_{j=1}^{i}(i < m)$, then $\alpha_R(D) < \delta_\alpha$ holds.

For the above example in Table 1, exclusive rule of m.c.h. is:

$$[prod = 0] \wedge [nau = 0] \wedge [M1 = 1] \to m.c.h. \quad \kappa = 1.0, \alpha = 1.0$$

This attains the highest accuracy. If the threshold for accuracy is set to 0.75, then

$$[prod = 0] \to m.c.h. \; \kappa = 1.0, \alpha = 0.75,$$
$$[nau = 0] \to m.c.h. \; \kappa = 1.0, \alpha = 0.75, \; and$$
$$[M1 = 1] \to m.c.h. \;\; \kappa = 1.0, \alpha = 1.0.$$

So, diagnostic rules for m.c.h. can be viewed as $1-$of-3 concept. In this way, combination of accuracy and coverage is also important to represent $m-$of$-n$ type rules.

5.2 Rule Induction Algorithm

An algorithm for induction of unit rules is derived by simple modification of the algorithm in Figure 1: if the thresholds of accuracy and coverage is set to δ and 1.0, respectively, then the algorithm for induction of each unit rule will be obtained. In this model, we should only add integration of unit rules after rule induction to obtain the total algorithm, which is not shown for the limitation of the space.

6 Conclusion

In this paper, rough set framework is introduced to model medical diagnostic rules. Acquired models show that the characteristics of medical reasoning reflect the concepts on approximation of rough sets, which explains why rough sets work well in medical domains.

References

1. Buchnan, B., Shortliffe, E.: Rule-Based Expert Systems. Addison-Wesley, New York (1984)
2. Tsumoto, S.: Automated extraction of medical expert system rules from clinical databases on rough set theory. Inf. Sci. 112, 67–84 (1998)
3. Tsumoto, S.: Extraction of experts decision rules from clinical databases using rough set model. Intelligent Data Analysis 2 (1998)
4. Pawlak, Z.: Rough Sets. Kluwer Academic Publishers, Dordrecht (1991)
5. Skowron, A., Grzymala-Busse, J.: From rough set theory to evidence theory. In: Yager, R., Fedrizzi, M., Kacprzyk, J. (eds.) Advances in the Dempster-Shafer Theory of Evidence, pp. 193–236. John Wiley & Sons, New York (1994)
6. Ziarko, W.: Variable precision rough set model. J. Comput. Syst. Sci. 46, 39–59 (1993)
7. Pawlak, Z.: Rough modus ponens. In: Proceedings of International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems 98, Paris (1998)
8. Tsumoto, S., Tanaka, H.: Automated knowledge acquisition from medical databases and its evaluation (1998)
9. Tsumoto, S., Tanaka, H.: Automated discovery of medical expert system rules from clinical databases based on rough sets. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining 96, Palo Alto, pp. 63–69. AAAI Press, California (1996)
10. Tsumoto, S.: Modelling medical diagnostic rules based on rough sets. In: Polkowski, L., Skowron, A. (eds.) RSCTC 1998. LNCS (LNAI), vol. 1424, pp. 475–482. Springer, Heidelberg (1998)
11. Langley, P.: Elements of Machine Learning. Morgan Kaufmann, San Francisco (1996)
12. McLachlan, G.: Discriminant Analysis and Statistical Pattern Recognition. John Wiley and Sons, New York (1992)

A Fundamentals of Accuracy and Coverage

A.1 Statistical Dependence

Let $P(R)$ and $P(D)$ be defined as: $P(R) = \frac{|R_A|}{|U|}$ *and* $P(D) = \frac{|D|}{|U|}$, where U denotes the total samples. Then, a index for statistical dependence ς_c is defined as:

$$\varsigma_R(D) = \frac{|R_A \cap D|}{|R_A||D|} = \frac{|U|P(R,D)}{P(R)P(D)},$$

where $P(R,D)$ denotes a joint probability of R and D ($P(R,D) = |R_A \cap D|/|U|$). Since the formula $P(R,D) = P(R)P(D)$ is the definition of statistical independence, $\varsigma_R(D)$ measures the degree of statistical dependence. That is, If $\varsigma_R(D) > 1.0$, then R and D are dependent, other R and D are independent; especially, if $\varsigma_R(D)$ is equal to 1.0, they are statistically independent.

Theorem 1. *Lower approximation and upper approximation gives (strong) statistical dependent relations.*

Proof. Since $\alpha_R(D) = 1.0$ *for the lower approximation,* $\varsigma_R(D) = \frac{1}{P(D)} > 1.0$ *In the same way, for the upper approximation,* $\varsigma_R(D) = \frac{1}{P(R)} > 1.0$ □

Definition 2. *Let U be described by n attributes. A conjunctive formula $R(i)$ is defined as: $R(i) = \bigwedge_{k=1}^{i}[a_i = v_i]$, where index i is sorted by a given criteria, such as the value of accuracy. Then, the sequence of a conjunction is given as: $R(i+1) = R(i) \wedge [a_{i+1} = v_{j+1}]$.*

Since $R(i+1)_A = R(i)_A \cap [a_{i+1} = v_{i+1}]_A$, for this sequence, the following proposition will hold: $R(i+1)_A \subseteq R(i)_A$ Thus, the following theorem is obtained.

Theorem 2. *When we consider a sequence of conjunctive formula such that the value of accuracy should be increased, the statistical dependence will increase.*
Proof.

$$\varsigma_{R(i+1)}(D) = \frac{\alpha_{R(i+1)}(D)}{P(D)} \geq \frac{\alpha_{R(i)}(D)}{P(D)} = \varsigma_{R(i)}(D)$$

A.2 Tradeoff Between Accuracy and Coverage

Theorem 3 (Monotonicity of Coverage). *Let a sequence of conjunctive formula $R(i)$ given with n attributes. Then,*

$$\kappa_{R(i+1)}(D) \leq \kappa_{R(i)}(D).$$

Then, since accuracy and coverage has the following relation:

$$\frac{\kappa_R(D)}{\alpha_R(D)} = \frac{P(R)}{P(D)}. \tag{1}$$

Since $P(R)$ will decrease with the sequence of conjunction, the following theorem is obtained.

Theorem 4. *Even if a sequence of conjunction for R is selected such that the value of accuracy increases monotonically, $\kappa_R(D)$ will decrease. That is, the decrease of $\kappa_R(D)$ is larger than the effect of the increase of $\alpha_R(D)$.* □