Computational Intelligence, Volume 17, Number 3, 2001

COMBINING ROUGH SETS AND BAYES' RULE

ZDZISŁAW PAWLAK
Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, ul. Bałtycka 5, 44 000 Gliwice, Poland

In rough set theory with every decision rule two conditional probabilities, called *certainty* and *coverage factors*, are associated. These two factors are closely related with the lower and the upper approximation of a set, basic notions of rough set theory. It is shown that these two factors satisfy the Bayes' rule.

The Bayes' rule in our case simply shows some relationship in the data, without referring to prior and posterior probabilities intrinsically associated with Bayesian inference. This relationship can be used to "invert" decision rules, i.e., to find reasons (explanation) for decisions thus providing inductive as well as deductive inference in our scheme.

Key words: Bayes' rule, rough sets, decision rules, information system.

1. INTRODUCTION

This paper contains an extended version of ideas presented in Pawlak (1998, 1999, 2000). In the paper relationship between rough set theory and Bayes' rule is shown. However, the meaning of Bayes' rule in our case is different from that in statistical inference.

Statistical inference grounded on the Bayes' rule supposes that some prior knowledge (prior probability) about some parameters, without knowledge about the data, is given first. Next the posterior probability is computed, when the data is available. The posterior probability is then used to verify the prior probability.

In the rough set philosophy with every decision rule two conditional probabilities, called *certainty* and *coverage factors*, are associated. These two factors are closely related with the lower and the upper approximation of a set, basic concepts of rough set theory. It turns out that these two factors satisfy the Bayes' rule. This property enables us to explain decisions in terms of conditions, i.e., to compute certainty and coverage factors of "inverse" decision rules, without referring to prior and posterior probabilities, intrinsically associated with Bayesian reasoning.

2. BASIC NOTIONS

In this section we recall basic concepts of rough set theory needed in the sequel. Starting point of rough set based data analysis is a data set, called an information system.

An information system is a data table, whose columns are labelled by attributes, rows are labelled by objects of interest and entries of the table are attribute values.

Formally, by an *information system* we will understand a pair $S = (U, A)$, where U and A are finite, nonempty sets called the *universe*, and the set of *attributes,* respectively. With every attribute $a \in A$ we associate a set V_a, of its *values*, called the *domain* of a. Any subset B of A determines a binary relation $I(B)$ on U, which will be called an *indiscernibility relation*, and defined as follows: $(x, y) \in I(B)$ if and only if $a(x) = a(y)$ for every $a \in A$, where $a(x)$ denotes the value of attribute a for element x.

Address correspondence to Z. Pawlak, at the Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, W. Bałtycka 5, 44 000 Gliwice, Poland; e-mail: zpw@ii.pw.edu.pl

Obviously $I(B)$ is an equivalence relation. The family of all equivalence classes of $I(B)$, i.e., a partition determined by B, will be denoted by $U/I(B)$, or simply by U/B; an equivalence class of $I(B)$, i.e., block of the partition U/B, containing x will be denoted by $B(x)$.

If (x, y) belongs to $I(B)$ we will say that x and y are *B-indiscernible* (*indiscernible with respect to B*). Equivalence classes of the relation $I(B)$ (or blocks of the partition U/B) are referred to as *B-elementary sets* or *B-granules*.

If we distinguish in an information system two disjoint classes of attributes, called *condition* and *decision attributes*, respectively, then the system will be called a *decision table* and will be denoted by $S = (U, C, D)$, where C and D are disjoint sets of condition and decision attributes, respectively.

A simple, tutorial example of an information system (a decision table) is shown in Table 1.

In Table 1 six facts concerning a hundred cases of driving a car in various weather conditions are presented. In the table *W (weather)*, *R (road)* and *T (time)* are *condition attributes*, representing driving conditions; *A (accident)*, is a *decision attribute*, giving information whether an accident has occurred or not. *N* is the number of similar cases. Each row of the decision table determines a decision obeyed when specified conditions are satisfied.

Suppose we are given an information system $S = (U, A)$, $X \subseteq U$, and $B \subseteq A$. Our task is to describe the set X in terms of attribute values from B. To this end we define two operations assigning to every $X \subseteq U$ two sets $B_*(X)$ and $B^*(X)$ called the *B-lower* and the *B-upper approximation* of X, respectively, and defined as follows:

$$B_*(X) = \bigcup_{x \in U} \{B(x) : B(x) \subseteq X\},$$

$$B^*(X) = \bigcup_{x \in U} \{B(x) : B(x) \cap X \neq \varnothing\}.$$

Hence, the B-lower approximation of a set is the union of all B-granules that are included in the set, whereas the *B-upper* approximation of a set is the union of all B-granules that have a nonempty intersection with the set. The set

$$BN_B(X) = B^*(X) - B_*(X)$$

will be referred to as the *B-boundary region* of X.

TABLE 1. An Example of an Information System

	driving conditions			*consequence*	
F	*W*	*R*	*T*	*A*	*N*
1	*misty*	*icy*	*day*	*yes*	6
2	*foggy*	*icy*	*night*	*yes*	8
3	*misty*	*not icy*	*night*	*yes*	5
4	*sunny*	*icy*	*day*	*no*	55
5	*foggy*	*not icy*	*dusk*	*yes*	11
6	*misty*	*not icy*	*night*	*no*	15

If the boundary region of X is the empty set, i.e., $BN_B(X) = \varnothing$, then X is *crisp* (*exact*) with respect to B; in the opposite case, i.e., if $BN_B(X) \neq \varnothing$, X is referred to as *rough* (*inexact*) with respect to B.

3. DECISION RULES

Let $S = (U, A)$ be an information system. With every $B \subseteq A$ we associate a formal language, i.e., a set of formulas F or (B). Formulas of F or (B) are built up from attribute-value pairs (a, v) where $a \in B$ and $v \in V_a$ by means of logical connectives $\wedge$ (*and*), $\vee$ (*or*), $\sim$ (*not*) in the standard way.

For any $\Phi \in F$ or (B) by $\|\Phi\|_S$ we denote the set of all objects $x \in U$ satisfying Φ in S and we refer to it as the *meaning* of Φ in S.

The meaning $\|\Phi\|_S$ of Φ in S is defined inductively as follows: $\|(a, v)\|_S = \{x \in U : a(v) = x\}$ for all $a \in B$ and $v \in V_a$, $\|\Phi \vee \Psi\|_S = \|\Phi\|_S \cup \|\Psi\|_S$, $\|\Phi \wedge \Psi\|_S = \|\Phi\|_S \cap \|\Psi\|_S$, $\| \sim \Phi\|_S = U - \|\Phi\|_S$.

A formula Φ is *true* in S if $\|\Phi\|_S = U$.

A *decision rule* in S is an expression $\Phi \rightarrow \Psi$, read *if* Φ *then* Ψ, where $\Phi \in F$ or (C), $\Psi \in F$ or (D) and C, D are condition and decision attributes, respectively; Φ and Ψ are referred to as *conditions* and *decisions* of the rule, respectively; we will also say that Φ is a condition for Ψ and Ψ is a decision for Φ.

The number $supp_S(\Phi, \Psi) = card(\|\Phi \wedge \Psi\|_S)$ will be referred to as the *support* of the decision rule $\Phi \rightarrow \Psi$ in S, whereas the number

$$\sigma_S(\Phi, \Psi) = \frac{supp_S(\Phi, \Psi)}{card(U)}$$

will be called the *strength* of the decision rule $\Phi \rightarrow \Psi$ in S.

A decision rule $\Phi \rightarrow \Psi$ is *true* in S if $\|\Phi\|_S \subseteq \|\Psi\|_S$.

We consider a probability distribution $p_U(x) = 1/card(U)$ for $x \in U$ where U is the (non-empty) universe of objects of S; we have $p_U(X) = card(X)/card(U)$ for $X \subseteq U$. For any formula Φ we associate its probability in S defined by

$$\pi_S(\Phi) = p_U(\|\Phi\|_S).$$

With every decision rule $\Phi \rightarrow \Psi$ we associate a conditional probability

$$\pi_S(\Psi|\Phi) = p_U(\|\Psi\|_S|\ \|\Phi\|_S)$$

that Ψ is true in S given Φ is true in S called the *certainty factor*, used first by Łukasiewicz (1970) to estimate the probability of implications. We have

$$\pi_S(\Psi|\Phi) = \frac{card(\|\Phi \wedge \Psi\|_S)}{card(\|\Phi\|_S)}$$

where $\|\Phi\|_S \neq \varnothing$.

This coefficient is now widely used in data mining and is called *confidence coefficient*.

Obviously, $\pi_S(\Psi|\Phi) = 1$ if and only if $\Phi \rightarrow \Psi$ is true in S.

If $\pi_S(\Psi|\Phi) = 1$, then $\Phi \rightarrow \Psi$ will be called a *certain decision* rule; if $0 < \pi_S(\Psi|\Phi) \leq 1$ the decision rule will be referred to as an *uncertain decision* rule.

Besides, we will also use a *coverage factor* (used e.g. by Tsumoto (1998) for estimation of the quality of decision rules) defined by

$$\pi_S(\Phi|\Psi) = p_U(\|\Phi\|_S|\ \|\Psi\|_S).$$

which is the conditional probability that Φ is true in S, given Ψ is true in S with the probability $\pi_S(\Psi)$. Obviously we have

$$\pi_S(\Phi|\Psi) = \frac{card(\|\Phi \wedge \Psi\|_S)}{card(\|\Psi\|_S)}.$$

The certainty factors in S can be also interpreted as frequencies of objects having the property Ψ in the set of objects having the property Φ and the coverage factors—as frequencies of objects having the property Φ in the set of objects having the property Ψ.

4. DECISION ALGORITHM

Let $Dec(S) = \{\Phi_i \to \Psi_i\}_{i=1}^m$, $m \geq 2$, be a set of decision rules in a decision table $S = (U, C, D)$.

If for every $\Phi \to \Psi$, $\Phi' \to \Psi' \in Dec(S)$ we have

(1) $\Phi = \Phi'$ or $\|\Phi \wedge \Phi'\|_S = \varnothing$,

and

(2) $\Psi = \Psi'$ or $\|\Psi \wedge \Psi'\|_S = \varnothing$,

then we will say that $Dec(S)$ is the set of *mutually disjoint* (*independent*) decision rules in S.

If

(3) $\| \bigvee_{i=1}^m \Phi_i\|_S = U$ and $\| \bigvee_{i=1}^m \Psi_i\|_S = U$

we will say that the set of decision rules $Dec(S)$ *covers* U.

If $\Phi \to \Psi \in Dec(S)$ and $\|\Phi \wedge \Psi\|_S \neq \varnothing$ we will say that the decision rule $\Phi \to \Psi$ is *admissible* in S.

If

(4) $\bigcup_{X \in U/D} C_*(X) = \| \bigvee_{\Phi \in Dec^+(S)} \Phi\|_S$

where $Dec^+(S)$ is the set of all certain decision rules in S, we will say that the set of decision rules $Dec(S)$ preserves the *consistency* of the decision table $S = (U, C, D)$.

Any set of decision rules $Dec(S)$ that is independent, covers U, preserves the consistency of S, and all decision rules $\Phi \to \Psi \in Dec(S)$ are admissible in S, will be called a *decision algorithm* in S.

An example of decision algorithm in S is given below:

1) $(W, misty) \wedge (R, icy) \to (A, yes)$
2) $(W, foggy) \to (A, yes)$
3) $(W, misty) \wedge (R, noticy) \to (A, yes)$
4) $(W, sunny) \to (A, no)$
5) $(W, misty) \wedge (R, not\ icy) \to (A, no)$

Hence, if $Dec(S)$ is a decision algorithm in S then the conditions of rules from $Dec(S)$ define in S a partition of U. Moreover, the *positive region of D with respect to C*, i.e., the set

$$\bigcup_{X \in U/D} C_*(X)$$

is partitioned by the conditions of some of these rules, which are certain in S.

TABLE 2. An Example of Certainty and Coverage Factors

rule no.	strength	certainty	coverage
1	0.06	1.00	0.20
2	0.19	1.00	0.63
3	0.05	0.25	0.17
4	0.55	1.00	0.79
5	0.15	0.75	0.21

Often we are interested in *explanation* of decisions in terms of conditions, i.e., in giving reasons to decisions. To this end, we have to "invert" decision rules, i.e., to exchange mutually conditions and decisions in a decision rule. The set of inverted decision rules will be called an *inverse decision algorithm*. For example, the following inverse decision algorithm can be understood as explanation of car accidents in terms of weather conditions:

1') $(A, yes) \rightarrow (R, icy) \wedge (W, misty)$
2') $(A, yes) \rightarrow (W, foggy)$
3') $(A, yes) \rightarrow (R, not\ icy) \wedge (W, misty)$
4') $(A, no) \rightarrow (W, sunny)$
5') $(A, no) \rightarrow (R, noticy) \wedge (W, misty)$

The certainty and the coverage factors for decision rules 1)–5) are shown in Table 2.

From decision rules 1)–5) and the certainty and the coverage factors we can draw the following conclusions:

1) Misty weather and icy road always caused accidents
2) Foggy weather always caused accidents
3) Misty weather and not icy road caused accidents *in 25% of cases*
4) Sunny weather and icy road always caused safe driving
5) Misty weather and not icy road caused safe driving in 75% of cases

From inverse decision rules 1')–5') we get the following explanations:

1') *20%* of accidents occurred when the weather was misty and the road icy
2') *63%* of accidents occurred when the weather was foggy
3') *17%* of accidents occurred when the weather was misty and the road not icy
4') *79%* of safe driving took place when the weather was sunny
5') *21%* of safe driving took place when the weather was misty and the road not icy

5. DECISION ALGORITHMS AND APPROXIMATIONS

There is an interesting relationship between decision algorithms and approximations, which is discussed below. Let $Dec(S)$ be a decision algorithm in S and let $\Phi \rightarrow \Psi \in Dec(S)$. By $C(\Psi)$ we will denote the set of all conditions of Ψ in $Dec(S)$ and by $D(\Phi)$—the set of all decisions of Φ in $Dec(S)$.

Then the following relationships are valid:

$$\text{a) } C_*(\|\Psi\|_S) = \left\| \bigvee_{\Phi' \in C(\Psi),\ \pi(\Psi|\Phi')=1} \Phi' \right\|_S,$$

$$\text{b) } C^*(\|\Psi\|_S) = \left\| \bigvee_{\Phi' \in C(\Psi),\ 0<\pi(\Psi|\Phi')\leq 1} \Phi' \right\|_S,$$

$$\text{c) } BN_C(\|\Psi\|_S) = \left\| \bigvee_{\Phi' \in C(\Psi),\ 0<\pi(\Psi|\Phi')<1} \Phi' \right\|_S.$$

The above properties enable us to introduce the following definitions:

i) If $\|\Phi\|_S = C_*(\|\Psi\|_S)$, then the formula Φ will be called the *C-lower approximation* of the formula Ψ and it will be denoted by $C_*(\Psi)$;
ii) If $\|\Phi\|_S = C^*(\|\Psi\|_S)$, then the formula Φ will be called the *C-upper approximation* of the formula Φ and it will be denoted by $C^*(\Psi)$;
iii) If $\|\Phi\|_S = BN_C(\|\Psi\|_S)$, then Φ will be called the *C-boundary* of the formula Ψ and it will be denoted by $BN_C(\Psi)$.

From the above considerations it follows that any decision $\Psi \in Dec(S)$ can be uniquely described by the following certain and uncertain decision rules respectively:

$$C_*(\Psi) \to \Psi,$$
$$BN_C(\Psi) \to \Psi.$$

Thus for the considered example we can get the following decision rules:

- Certain rules describing accidents (the lower approximation of the set of facts $\{1, 2, 3, 5\}$)

  1) $(W, misty) \wedge (R, icy) \to (A, yes)$
  2) $(W, foggy) \to (A, yes)$

- Uncertain rule describing accidents (the boundary region $\{3, 6\}$ of the set of facts $\{1, 2, 3, 5\}$)

  3) $(W, misty) \wedge (R, not\ icy) \to (A, yes)$

- Certain rule describing lack of accidents (the lower approximation of the set of facts $\{4, 6\}$)

  4) $(W, sunny) \to (A, no)$

- Uncertain rule describing lack of accidents (the boundary region $\{3, 6\}$ of the set of facts $\{4, 6\}$)

  5) $(W, misty) \wedge (R, not\ icy) \to (A, no)$

This coincides with the idea given by Ziarko (1998) to represent decision tables by means of three types of decision rules corresponding to the positive region, the boundary region, and the negative region of a decision.

6. SOME PROPERTIES OF DECISION ALGORITHMS

Let $Dec(S)$ be a decision algorithm and let $\Phi \to \Psi \in Dec(S)$. Then the following properties are valid:

$$\sum_{\Phi' \in C(\Psi)} \pi_S(\Phi'|\Psi) = 1 \tag{1}$$

$$\sum_{\Psi' \in D(\Phi)} \pi_S(\Psi'|\Phi) = 1 \tag{2}$$

$$\pi_S(\Psi) = \sum_{\Phi' \in C(\Psi)} \pi_S(\Psi|\Phi') \cdot \pi_S(\Phi') = \sum_{\Phi' \in C(\Psi)} \sigma_S(\Phi', \Psi) \tag{3}$$

$$\pi_S(\Phi|\Psi) = \frac{\pi_S(\Psi|\Phi) \cdot \pi_S(\Phi)}{\sum_{\Phi' \in C(\Psi)} \pi_S(\Psi|\Phi') \cdot \pi_S(\Phi')} = \frac{\sigma_S(\Phi, \Psi)}{\pi_S(\Psi)} \tag{4}$$

That is, any decision algorithm satisfies (1), (2), (3), and (4). Let us observe that (3) is the well known *total probability* formula and (4) is the Bayes' rule.

The Bayes' rule explains the relationships between the certainty and coverage factors and can be used to explain decisions in terms of conditions.

The Bayes' rule (4) enables us to compute in a very simple way, employing only the strength of decision rules, the certainty and the coverage factors.

Let us notice that we do not refer to prior and posterior probabilities, inherently associated with Bayesian inference. In our case the Bayes' rule simply reveals some relationships in the data.

7. CONCLUSIONS

This paper shows that any decision algorithm is closely associated with approximations, moreover it satisfies Bayes' rule. This enables us to apply Bayes' rule to "invert" decision rules without referring to prior and posterior probabilities, inherently associated with "classical" Bayesian inference philosophy, using only strength of decision rules. The inverse decision rules can be used to explain decisions in terms of conditions, i.e., to give reasons for decisions.

ACKNOWLEDGMENTS

Thanks are due to Professor Andrzej Skowron and to the anonymous referee for their critical remarks.

REFERENCES

ADAMS, E. W. 1975. The logic of conditionals, an application of probability to deductive logic. D. Reidel Publishing Company, Dordrecht, Boston.

GRZYMAŁA-BUSSE, J. 1991. Managing Uncertainty in Expert Systems. Kluwer Academic Publishers, Boston, Dordrecht.

ŁUKASIEWICZ, J. 1970. Die logishen Grundlagen der Wahrscheinilchkeitsrechnung. Krakow, 1913. *In* Jan Łukasiewicz—Selected Works. *Edited by* L. Borkowski. North Holland Publishing Company, Amsterdam, London, Polish Scientific Publishers, Warsaw.

PAWLAK, Z. 1991. Rough Sets—Theoretical Aspects of Reasoning about Data; Kluwer Academic Publishers: Boston, Dordrecht.

PAWLAK, Z. 1998. Reasoning about data—a rough set perspective. *In* Proceedings of the First International Conference Rough Sets and Current Trends in Computing (RSCTC'98), *Edited by* L. Polkowski and A. Skowron. Warsaw, Poland, June, Lecture Notes in Artificial Intelligence 1424 Springer-Verlag, Berlin, pp. 25–34.

PAWLAK, Z. 1999. Decision Rules, Bayes' Rule and Rough Sets. *In* Proceedings of 7th International Workshop: New Directions in Rough Sets, Data Mining, and Granular—Soft Computing (RSFDGSC'99). *Edited by* N. Zhong, A. Skowron, S. Ohsuga. Yamaguchi, Japan, November, Lecture Notes in Artificial Intelligence 1711 Springer-Verlag, Berlin, pp. 1–9.

PAWLAK, Z. 2000. Drawing Conclusions from Data—the Rough Set Way, May 29, 2001.

PAWLAK, Z, and A. SKOWRON, 1994. Rough membership functions. *In* Advances in the Dempster Shafer Theory of Evidence. *Edited by* R. R. Yaeger, M. Fedrizzi, and J. Kacprzyk. John Wiley & Sons, Inc., New York, pp. 251–271.

POLKOWSKI, L., and A. SKOWRON, Eds. 1998. Proceedings of the First International Conference Rough Sets and Current Trends in Computing (RSCTC'98), Warsaw, Poland, June, Lecture Notes in Artificial Intelligence 1424, Springer-Verlag, Berlin.

POLKOWSKI, L., and A. SKOWRON, Eds. 1998. Rough Sets in Knowledge Discovery Vol. 1–2, Physica-Verlag, Heidelberg.

SKOWRON, A. 1994. Menagement of uncertainty in AI: A rough set approach. *In* Proceedings of the conference SOFTEKS, Springer-Verlag and British Computer Society, pp. 69–86.

TSUMOTO, S., S. KOBAYASHI, T. YOKOMORI, H. TANAKA, and A. NAKAMURA, Eds. 1996. Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery (RSFD'96). The University of Tokyo, November 6–8.

TSUMOTO, S. 1998. Modelling medical diagnostic rules based on rough sets. *In* Proceedings of the First International Conference Rough Sets and Current Trends in Computing (RSCTC'98). *Edited by* L. Polkowski, A. Skowron. Warsaw, Poland, June, Lecture Notes in Artificial Intelligence 1424, Springer-Verlag, Berlin, pp. 475–482.

ZIARKO, W. 1998. Approximation Region-Based Decision Tables. *In* Rough Sets in Knowledge Discovery Vol. 1–2. *Edited by* L. Polkowski, A. Skowron. Physica-Verlag, Heidelberg, pp. 178–185.