Drawing Conclusions from Data — The Rough Set Way

Zdzisław Pawlak*
Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, ul. Bałtycka 5, 44 000 Gliwice, Poland

In the 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' theorem. The Bayes' theorem in our case simply shows some relationship in the data, without referring to prior and posterior probabilities intrinsically associated with Bayesian inference in our case and can be used to "inverse" decision rules, i.e., to find reasons (explanation) for decisions. © 2001 John Wiley & Sons, Inc.

1. INTRODUCTION

This paper is a modified version of Ref. 1. The relationship between rough set theory (see Refs. 7–11) and Bayes' theorem is shown. However the meaning of Bayes' theorem in our case and that in statistical inference is different.

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 turned out that these two factors satisfy the Bayes' theorem without referring to prior and posterior probabilities. 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.

* e-mail: zpw@ii.pw.edu.pl.

INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS, VOL. 16, 3–11 (2001)

2. INFORMATION SYSTEM AND DECISION TABLE

The 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 labeled by attributes, rows are labeled 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 is 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. Obviously $I(B)$ is an equivalence relation. The family of all equivalence classes of $I(B)$, i.e., partition determined by B, will be denoted by $U/I(B)$, or simple 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* or 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 classes of attributes, called *condition* and *decision attributes*, respectively, then the system will be called a *decision table*.

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

The table contains data about six car types, where F, P, and S are condition attributes and denote *fuel consumption*, *selling price* and *size* respectively, whereas M denotes marketability and is the decision attribute.

Besides, T denotes the car type and N the number of cars sold of a given type.

Each row of the decision table determines a decision obeyed when specified conditions are satisfied.

Table I. An example of an information system.

T	F	P	S	M	N
1	med.	med.	med.	poor	8
2	high	med.	large	poor	10
3	med.	low	large	poor	4
4	low	med.	med.	good	50
5	high	low	small	poor	8
6	med.	low	large	good	20

3. APPROXIMATIONS

Suppose we are given an information system (a data set) $S = (U, A)$, a subset X of the universe U, and subset of attributes B. 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:

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

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

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.

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

For example, for $B = \{F, P, S\}$ and the set $X = \{[1] \cup [2] \cup [3] \cup [5]\}$ of cars with poor marketability we have $B_*(X) = \{[1] \cup [2] \cup [5]\}$, $B^*(X) = \{[1] \cup [2] \cup [3] \cup [5] \cup [6]\}$ and $BN_B(X) = \{[3] \cup [6]\}$, where $[i]$ denotes the set of all cars of type i.

4. DECISION RULES

With every information system $S = (U, A)$ we associate a formal language $L(S)$, written L when S is understood. Expressions of the language L are logical formulas denoted by Φ, Ψ, etc. built up from attributes and attribute-value pairs by means of logical connectives $\wedge$ (*and*), $\vee$ (*or*), $\sim$ (*not*) in the standard way. We will denote by $\|\Phi\|_S$ the set of all objects $x \in U$ satisfying Φ in S and refer to as the *meaning* of Φ in S.

The meaning of Φ in S is defined inductively as follows:

(1) $\|(a, v)\|_S = \{x \in U : a(v) = x\}$ for all $a \in A$ and $v \in V_a$
(2) $\|\Phi \vee \Psi\|_S = \|\Phi\|_S \cup \|\Psi\|_S$
(3) $\|\Phi \wedge \Psi\|_S = \|\Phi\|_S \cap \|\Psi\|_S$
(4) $\| \sim \Phi\|_S = U - \|\Phi\|_S$.

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

A *decision rule* in L is an expression $\Phi \rightarrow \Psi$, read *if* Ψ *then* Ψ; Φ and Ψ are referred to as *conditions* and *decisions* of the rule, respectively.

An example of a decision rule is given below:

$$(F, \text{med}) \wedge (P, \text{low}) \wedge (S, \text{large}) \rightarrow (M, \text{poor})$$

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

With every decision rule $\Phi \rightarrow \Psi$ we associate a conditional probability $\pi_S(\Psi \mid \Phi)$ that Ψ is true in S given Φ is true in S with the probability $\pi_S(\Phi) = \text{card}(\|\Phi\|_S)/\text{card}(U)$, called the *certainty factor* introduced first by Łukasiewicz[2] (see also 5, 6, and 12) and defined as follows:

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

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

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

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

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

Besides, we will also use a *coverage factor* proposed by Tsumoto[3,4] and defined as follows:

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

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

5. DECISION RULES AND APPROXIMATIONS

Let $\{\Phi_i \rightarrow \Psi\}_n$ be a set of n decision rules $\Phi_1 \rightarrow \Psi, \Phi_2 \rightarrow \Psi, \ldots, \Phi_n \rightarrow \Psi$ such that:

$$\begin{aligned} &\text{all conditions } \Phi_i \text{ are pairwise mutually exclusive, i.e., } \|\Phi_i \wedge \Phi_j\|_S = \varnothing, \\ &\text{for any } 1 \leq i, j \leq n, i \neq j, \text{ and } \sum_{i=1}^{n} \pi_S(\Phi_i \mid \Psi) = 1. \end{aligned} \tag{1}$$

Let C be the set of condition attributes and let $\{\Phi_i \rightarrow \Psi\}_n$ be a set of decision rules satisfying eq. (1).

Then the following relationships are valid:

(a) $C_*(\|\Psi\|)_S = \left\| \bigvee_{\pi(\Psi \mid \Phi_i) = 1} \Phi_i \right\|_S$

(b) $C^*(\|\Psi\|_S) = \left\| \bigvee_{0 < \pi(\Psi \mid \Phi_i) \leq 1} \Phi_i \right\|_S$

(c) $BN_C(\|\Psi\|)_S = \left\| \bigvee_{0<\pi(\Psi|\Phi_i)<1} \Phi_i \right\|_S$.

The above properties enable us to introduce the following definitions:

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

In this way we are allowed to approximate not only sets but also formulas.

Let us consider the following example.

The C-lower approximation of $(M, poor)$ is the formula

$$\begin{aligned} C_*(M, poor) = &((F, med.) \wedge (P, med.) \wedge (S, med.)) \\ &\vee ((F, high) \wedge (P, med.) \wedge (S, large)) \\ &\vee ((F, high) \wedge (P, low) \wedge (S, small)) \end{aligned}$$

The C-upper approximation of $(M, poor)$ is the formula

$$\begin{aligned} C^*(M, poor) = &((F, med.) \wedge (P, med.) \wedge (S, med.)) \\ &\vee ((F, high) \wedge (P, med.) \wedge (S, large)) \\ &\vee ((F, med.) \wedge (P, low) \wedge (S, large)) \\ &\vee ((F, high) \wedge (P, low) \wedge (S, small)) \end{aligned}$$

The C-boundary of $(M, poor)$ is the formula

$$BN_C(M, poor) = (F, med.) \wedge (P, low) \wedge (S, large)$$

After simplification using the rough set approach (not presented here) we get the following approximations:

$$\begin{aligned} C_*(M, poor) &= ((F, med.) \wedge (P, med.)) \vee (F, high) \\ C^*(M, poor) &= (F, med.) \vee (F, high) \end{aligned}$$

From the above considerations it follows that any decision Ψ can be uniquely described by the following decision rules:

$$\begin{aligned} C_*(\Psi) &\to \Psi, \\ C^*(\Psi) &\to \Psi, \\ BN_C(\Psi) &\to \Psi, \end{aligned}$$

or equivalently

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

$$BN_C(\Psi) \to \Psi.$$

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

$$((F, med.) \wedge (P, med.)) \vee (F, high) \to (M, poor)$$

$$(F, med.) \wedge (P, low) \to (M, poor)$$

which are associated with the lower approximation and the boundary region of the decision $(M, poor)$, respectively and describe decision $(M, poor)$.

For the decision $(M, good)$ we get the following decision rules:

$$(F, low) \to (M, good)$$

$$(F, med.) \wedge (P, low) \to (M, good)$$

This coincides with the idea given by Ziarko[13] to represent decision tables by means of three decision rules corresponding to positive region, the boundary region, and the negative region of a decision.

6. INVERSE DECISION RULES

Often we are interested in *explanation* of decisions in terms of conditions, i.e., give reasons to decisions. To this end we have to "inverse" the decision rules, i.e., to exchange mutually conditions and decisions in a decision rule.

For example, for the set of decision rules describing poor and good marketability:

	cer.	*cov.*	
$((F, med.) \wedge (P, med.)) \vee (F, high) \to (M, poor)$	1.00	0.87	
$((F, med.) \wedge (P, low) \to (M, poor)$	0.17	0.13	(1)
$(F, low) \to (M, good)$	1.00	0.71	
$(F, med.) \wedge (P, low) \to (M, good)$	0.83	0.29	

we get the following inverse decision rules:

	cer.	*cov.*	
$(M, poor) \to ((F, med.) \wedge (P, med.)) \vee (F, high)$	0.87	1.00	
$(M, poor) \to ((F, med.) \wedge (P, low))$	0.13	0.17	(2)
$(M, good) \to (F, low)$	0.71	1.00	
$(M, good) \to (F, med.) \wedge (P, low)$	0.83	0.83	

We can get more specific decision rules replacing sets of decision rules (1) and (2) by means of simple decision rules (a decision rule is simple if it does not contain logical connectives $\vee$).

Thus instead of decision rules (1) we will have:

$$
\begin{array}{lll}
 & cer. & cov. \\
(F, med.) \wedge (P, med.) \rightarrow (M, poor) & 1.00 & 0.27 \\
(F, high) \rightarrow (M, poor) & 1.00 & 0.60 \\
(F, med.) \wedge (P, low) \rightarrow (M, poor) & 0.17 & 0.13 \\
(F, low) \rightarrow (M, good) & 1.00 & 0.71 \\
(F, med.) \wedge (P, low) \rightarrow (M, good) & 0.83 & 0.29
\end{array} \tag{3}
$$

and decision rules (2) can be replaced by:

$$
\begin{array}{lll}
 & cer. & cov. \\
(M, poor) \rightarrow (F, med.) \wedge (P, med.) & 0.27 & 1.00 \\
(M, poor) \rightarrow (F, high) & 0.60 & 1.00 \\
(M, poor) \rightarrow (F, med.) \wedge (P, low) & 0.13 & 0.17 \\
(M, good) \rightarrow (F, low) & 0.71 & 1.00 \\
(M, good) \rightarrow (F, med.) \wedge (P, low) & 0.29 & 0.89
\end{array} \tag{4}
$$

From the set of decision rules and their certainty and coverage factors we can draw the following conclusions:

(1) Cars with medium fuel consumption and medium price or high fuel consumption always have poor marketability (sell poorly).
(2) 17% cars with medium fuel consumption and low price have poor marketability (sell poorly).
(3) Cars with low fuel consumption always have good marketability (sell well).
(4) 83% cars with medium fuel consumption and low price have good marketability (sell well).

The above conclusions can be explained by means of inverse decision rules as follows:

(1′) 87% cars selling poorly have medium fuel consumption and medium price (27%) or high fuel consumption (60%).
(2′) 13% cars selling poorly have medium fuel consumption and low price.
(3′) 71% cars with low fuel consumption are selling well.
(4′) 83% cars selling well have medium fuel consumption and low price.

7. PROPERTIES OF DECISION RULES

If $\{\Phi_i \rightarrow \Psi\}_n$ is a set of decision rules satisfying condition (1), then the well known formula for total probability holds:

$$\pi_S(\Psi) = \sum_{i=1}^{n} \pi_S(\Psi \mid \Phi_i) \cdot \pi_S(\Phi_i) \tag{2}$$

Moreover for any decision rule $\Phi \to \Psi$ the following Bayes' theorem is valid:

$$\pi_S(\Phi_j \mid \Psi) = \frac{\pi_S(\Psi \mid \Phi_j) \cdot \pi_S(\Phi_j)}{\sum_{i=1}^{n} \pi_S(\Psi \mid \Phi_i) \cdot \pi_S(\Phi_i)} \tag{3}$$

That is, any decision table or any set of implications satisfying condition (1) satisfies the Bayes' theorem, without referring to prior and posterior probabilities—fundamental in Bayesian data analysis philosophy. Bayes' theorem in our case says that: if an implication $\Phi \to \Psi$ is true to the degree $\pi_S(\Psi \mid \Phi)$ then the implication $\Psi \to \Phi$ is true to the degree $\pi_S(\Phi \mid \Psi)$.

This property explains the relationships between the certainty and coverage factors and can be used to explain decisions in terms of conditions, i.e., it can be used to compute coverage factors by means of certainty factors, but this is more complicated than the direction computation from the data. The Bayes' theorem is more useful when instead of data table like Table I we are given some probabilities, but we will not discuss this issue in this paper.

8. CONCLUSIONS

It is shown in this paper that any decision table satisfies Bayes' theorem. This enables us to apply Bayes' theorem to "inverse" decision rules without referring to prior and posterior probabilities, inherently associated with "classical" Bayesian inference philosophy.

The inverse decision rules can be used to explain decisions in terms of conditions, i.e., to give reasons for decisions.

Let us observe that the conclusions drawn from the data are not universally true, but are valid only for the data. Whether they are valid for a bigger universe depends if the data is a proper sample of the bigger universe or not.

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

References

1. Pawlak Z. Decision rules, Bayes' rule and rough sets. In: Zhong N, Skoron A, Ohsuga S, editors. New directions in rough sets, data mining, and granular—soft computing. Proceedings 7th International Workshop, RSFDGSC'99, Yamaguchi, Japan, November 1999. p 1–9.
2. Łukasiewicz J. Die logishen Grundlagen der Wahrscheinilchkeitsrechnung. Krakow (1913). In: Borkowski L, editor. Jan Łukasiewicz—Selected works. Amsterdam, London: North Holland Publishing; Warsaw: Polish Scientific Publishers; 1970.
3. Tsumoto S, Kobayashi S, Yokomori T, Tanaka H, Nakamura A, editors. Proceedings of the fourth international workshop on rough sets, fuzzy sets, and machine discovery (RSFD'96). The University of Tokyo, November 6–8, 1996.
4. Tsumoto S. Modelling medical diagnostic rules based on rough sets. In: Polkowski L, Skowron A, editors. Rough sets and current trends in computing. Lecture notes in artificial intelligence. Proceedings, First International Conference, RSCTC'98, Warsaw, Poland, June, 1998. p 475–482.

5. Adams, EW. The logic of conditionals, an application of probability to deductive logic. Boston, Dordrecht: D Reidel Publishing Company; 1975.
6. Grzymała–Busse J. Managing uncertainty in expert systems. Boston, Dordrecht: Kluwer Academic Publishers; 1991.
7. Pawlak Z. Rough sets—theoretical aspects of reasoning about data. Boston, Dordrecht: Kluwer Academic Publishers; 1991.
8. Pawlak Z. Reasoning about data—a rough set perspective. In: Polkowski L, Skowron, A, editors. Rough sets and current trends in computing. Lecture notes in artificial intelligence. First international conference, RSCTC'98, Warsaw, Poland, June 1998. p 25–34.
9. Pawlak Z, Skowron A. Rough membership functions. In: Yaeger RR, Fedrizzi M, Kacprzyk J, editors. Advances in the Dempster Shafer theory of evidence. New York: John Wiley & Sons, Inc.; 1994. p 251–271.
10. Polkowski L, Skowron A, editors. Rough sets and current trends in computing. Lecture notes in artificial intelligence. Proceedings First International Conference, RSCTC'98, Warsaw, Poland, June, 1998.
11. Polkowski L, Skowron A, editors. Rough sets in knowledge discovery. Physica-Verlag, Vol. 1, No. 2, 1998.
12. Skowron A. Management of uncertainty in AI: a rough set approach. In: Proceedings of the Conference SOFTEKS. Springer Verlag and British Computer Society; 1994. p 69–86.
13. Ziarko W. Approximation region-based decision tables. In: Polkowski L, Skowron A, editors. Rough sets in knowledge discovery. Physica-Verlag, Vol. 1, No. 2, 1998. p 178–185.