Bayes' Theorem - the Rough Set Perspective

Zdzisław Pawlak

Institute of Theoretical and Applied Informatics, Polish Academy of Sciences,
ul. Bałtycka 5, 44 100 Gliwice, Poland

MOTTO:
"I had come to an entirely erroneus conclusions,
which shows, my dear Watson, how dangerous
it always is to reason from insufficient data"
Sherlock Holmes
In: "The speckled band"

Abstract. Rough set theory offers new insight into Bayes' theorem. It does not refer either to prior or posterior probabilities, inherently associated with Bayesian reasoning, but reveals some probabilistic structure of the data being analyzed. This property can be used directly to draw conclusions from data.

It is also worth mentioning the relationship between Bayes' theorem and flow graphs.

1 Introduction

This article is a modified version of paper [9].

Bayes' theorem is the essence of statistical inference.

Bayes formulated the following problem: "*Given* the number of times in which an unknown event has happened and failed: *required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named" [2].

In spite of great power of Bayesian inference process the method also has caused many criticism, as it can be seen, e.g., from the following excerpts.

"The technical results at the heart of the essay is what we now know as *Bayes' theorem*. However, from a purely formal perspective there is no obvious reason why this essentially trivial probability result should continue to excite interest" [3].

"Opinion as to the values of Bayes' theorem as a basic for statistical inference has swung between acceptance and rejection since its publication on 1763" [4].

In fact "... it was Laplace (1774 – 1886) – apperently unaware of Bayes' work – who stated the theorem in its general (discrete) from" [3].

Rough set theory throws a new light on Bayes' theorem. The proposed approach does not refer either to prior or posterior probabilities, inherently

associated with Bayesian reasoning, but it reveals some probabilistic structure of the data being analyzed, i.e., it states that any data set (decision table) satisfies total probability theorem and Bayes' theorem, which can be used directly to draw conclusions from data.

The rough set approach to Bayes' theorem shows close relationship between logic of implications and probability, which was first observed by Łukasiewicz [7] and also independly studied by Adams [1] and others. Bayes' theorem in this context can be used to "invert" implications, i.e. to give reasons for decisions.

Besides, we propose a new form of Bayes' theorem where basic role is played by strength of decision rules (implications) derived from the data. The strength of decision rules is computed from the data or it can be also a subjective assessment. This formulation gives new look on Bayesian methodology of inference and also essentially simplifies computations.

It is also worth mentioning the relationship between Bayes' theorem and flow graphs, which leads to a new kind of flow networks, different than those of Ford and Fulkerson [6].

2 Bayes' Theorem

In this section we recall basic ideas of Bayesian inference philosophy, after [3–5].

"In its simplest form, if H denotes an hypothesis and D denotes data, the theorem says that

$$P\left(H|D\right) = P\left(D|H\right) \times P\left(H\right) / P\left(D\right).$$

With $P(H)$ regarded as a probabilistic statement of belief about H before obtaining data D, the left-hand side $P(H|D)$ becomes an probabilistic statement of belief about H after obtaining D. Having specified $P(D|H)$ and $P(D)$, the mechanism of the theorem provides a solution to the problem of how to learn from data.

In this expression, $P(H)$, which tells us what is known about H without knowing of the data, is called the *prior* distribution of H, or the distribution of H *a priori*. Correspondingly, $P(H|D)$, which tells us what is known about H given knowledge of the data, is called the *posterior* distribution of H given D, or the distribution of H a *posteriori*" [3].

"A prior distribution, which is supposed to represent what is known about unknown parameters before the data is available, plays an important role in Baysian analysis. Such a distribution can be used to represent prior knowledge or relative ignorance" [4].

Let us illustrate the above by a simple example taken from [5].

Example 1. "Consider a physician's diagnostic test for presence or absence of some rare disease D, that only occurs in 0.1% of the population, i.e.,

$P(D) = .001$. It follows that $P(\overline{D}) = .999$, where $\overline{D}$ indicates that a person does not have the disease. The probability of an event before the evaluation of evidence through Bayes' rule is often called the prior probability. The prior probability that someone picked at random from the population has the disease is therefore $P(D) = .001$.

Furthermore we denote a positive test result by T^+, and a negative test result by T^-. The performance of the test is summarized in Table 1.

Table 1. Performance of diagnostic test

	T^+	T^-
D	0.95	0.05
$\overline{D}$	0.02	0.98

What is the probability that a patient has the disease, if the test result is positive? First, notice that $D, \overline{D}$ is a partition of the outcome space. We apply Bayes' rule to obtain

$$P\left(D|T^+\right) = \frac{P\left(T^+|D\right)P\left(D\right)}{P\left(T^+|D\right)P\left(D\right) + P\left(T^+|\overline{D}\right)P\left(\overline{D}\right)} = \\ = \frac{.95 \cdot .001}{.95 \cdot .001 + .02 \cdot .999} = .045.$$

Only 4.5% of the people with a positive test result actually have the disease. On the other hand, the posterior probability (i.e. the probability after evaluation of evidence) is 45 times as high as the prior probability".

3 Information Systems and Approximation of Sets

In this section we define basic concepts of rough set theory: information system and approximation of sets. Rudiments of rough set theory can be found in [8,11].

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 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., 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.

Thus the decision table determines decisions which must be taken, when some conditions are satisfied. In other words each row of the decision table specifies a decision rule which determines decisions in terms of conditions.

Observe, that elements of the universe are in the case of decision tables simply labels of decision rules.

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:

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

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

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.

4 Rough Membership

Rough sets can be also defined employing instead of approximations rough membership function [10], which is defined as follows:

$$\mu_X^B : U \to [0, 1]$$

and

$$\mu_X^B(x) = \frac{|B(x) \cap X|}{|B(x)|},$$

where $X \subseteq U$, $B \subseteq A$ and $|X|$ denotes the cardinality of X.

The function measures the degree that x belongs to X in view of information about x expressed by the set of attributes B.

It can be shown [10] that the rough membership function has the following properties:

1. $\mu_X^B(x) = 1$ *iff* $x \in B_*(X)$
2. $\mu_X^B(x) = 0$ *iff* $x \in U - B^*(X)$
3. $0 < \mu_X^B(x) < 1$ *iff* $x \in BN_B(X)$
4. $\mu_{U-X}^B(x) = 1 - \mu_X^B(x)$ for any $x \in U$
5. $\mu_{X \cup Y}^B(x) \geq max\left(\mu_X^B(x),\ \mu_Y^B(x)\right)$ for any $x \in U$
6. $\mu_{X \cap Y}^B(x) \leq min\left(\mu_X^B(x),\ \mu_Y^B(x)\right)$ for any $x \in U$

Observe that the rough membership function is a generalization of fuzzy membership function since properties 5) and 6) are more general than the corresponding properties for fuzzy membership.

5 Decision Language

It is often useful to describe decision tables in logical terms. To this end we associate with every decision table $S = (U, C, D)$ a formal language called a *decision language* denoted $L(S)$.

Let $S = (U, A)$ be a decision table. With every $B \subseteq A = C \cup D$ we associate a set of formulas $For(B)$. Formulas of $For(B)$ 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 $\Phi \in For(B)$ by $\|\Phi\|_S$ we denote the set of all objects $x \in U$ satisfying Φ in S 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 *decision rule* in $L(S)$ is an expression $\Phi \rightarrow_S \Psi$, or simply $\Phi \rightarrow \Psi$ if S is understood, read *if* Φ *then* Ψ, where $\Phi \in For(C)$, $\Psi \in For(D)$ and C, D are condition and decision attributes, respectively; Φ and Ψ are referred to as *conditions* part and *decisions* part of the rule, respectively.

The number $supp_S(\Phi, \Psi) = |(||\Phi \wedge \Psi||_S)|$ will be called the *support* of the rule $\Phi \rightarrow \Psi$ in S. We consider a probability distribution $p_U(x) = 1/|U|$ for $x \in U$ where U is the (nonempty) universe of objects of S; we have $p_U(X) = |X|/|U|$ 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 $\Phi \to \Psi$ we associate a conditional probability

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

called the *certainty factor* of the decision rule, denoted $cer_S(\Phi, \Psi)$. This idea was used first by Łukasiewicz [7] (see also [1]) to estimate the probability of implications. We have

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

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

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

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

There is an interesting relationship between decision rules and their approximations: certain decision rules correspond to the lower approximation, whereas the uncertain decision rules correspond to the boundary region.

Besides, we will also use a *coverage factor* of the decision rule, denoted $cov_S(\Phi, \Psi)$ (used e.g., by Tsumoto and Tanaka [12] for estimation of the quality of decision rules) defined by

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

Obviously we have

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

There are several possibilities to interpret the certainty and the coverage factors: statistical (frequency), probabilistic (conditional probability), logical (degree of truth), mereological (degree of inclusion) and set theoretical (degree of membership).

We will use here mainly the statistical interpretation, i.e., the certainty factors will be interpreted as the frequency of objects having the property Ψ in the set of objects having the property Φ and the coverage factor – as the frequency of objects having the property Φ in the set of objects having the property Ψ.

Let us observe that the factors are not assumed arbitrarily but are computed from data.

The number

$$\sigma_S(\Phi, \Psi) = \frac{supp_S(\Phi, \Psi)}{|U|} = \pi_S(\Psi|\Phi) \cdot \pi_S(\Phi)$$

will be called the *strength* of the decision rule $\Phi \to \Psi$ in S, and will play an important role in our approach.

We will need also the notion of equivalence of formulas.

Let Φ, Ψ be formulas in $For(A)$ where A is the set of attributes in $S = (U, A)$.

We say that Φ and Ψ are *equivalent* in S, or simply, equivalent if S is understood, in symbols $\Phi \equiv \Psi$, if and only if $\Phi \to \Psi$ and $\Psi \to \Phi$. It means that $\Phi \equiv \Psi$ if and only if $||\Phi||_S = ||\Psi||_S$.

We need also *approximate equivalence* of formulas which is defined as follows:

$$\Phi \equiv_k \Psi \text{ if and only if } cer(\Phi, \Psi) = cov(\Phi, \Psi) = k.$$

Besides, we define also *approximate equivalence* of formulas with the *accuracy* $\varepsilon\, (0 \leq \varepsilon \leq 1)$, which is defined as follows:

$$\Phi \equiv_{k,\varepsilon} \Psi \text{ if and only if } k = min\{cer(\Phi, \Psi), cov(\Phi, \Psi)\}$$

$$\text{and } |cer(\Phi, \Psi) - cov(\Phi, \Psi)| \leq \varepsilon.$$

6 Decision Algorithms

In this section we define the notion of a decision algorithm, which is a logical counterpart of a decision table.

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

1) If for every $\Phi \to \Psi$, $\Phi' \to \Psi' \in Dec(S)$ we have $\Phi = \Phi'$ or $\|\Phi \wedge \Phi'\|_S = \emptyset$, and $\Psi = \Psi'$ or $\|\Psi \wedge \Psi'\|_S = \emptyset$, then we will say that $Dec(S)$ is the set of pairwise *mutually exclusive* (*independent*) decision rules in S.
2) If $\left\| \bigvee_{i=1}^m \Phi_i \right\|_S = U$ and $\left\| \bigvee_{i=1}^m \Psi_i \right\|_S = U$ we will say that the set of decision rules $Dec(S)$ *covers* U.
3) If $\Phi \to \Psi \in Dec(S)$ and $supp_S(\Phi, \Psi) \neq 0$ we will say that the decision rule $\Phi \to \Psi$ is *admissible* in S.
4) If $\bigcup_{X \in U/D} C_*(X) = \left\| \bigvee_{\Phi \to \Psi \in Dec^+(S)} \Phi \right\|_S$, where $Dec^+(S)$ is the set of all certain decision rules from $Dec(S)$, we will say that the set of decision rules $Dec(S)$ preserves the *consistency* part of the decision table $S = (U, C, D)$.

The set of decision rules $Dec(S)$ that satisfies 1), 2) 3) and 4), i.e., 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.

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.

If $\Phi \to \Psi$ is a decision rule then the decision rule $\Psi \to \Phi$ will be called an *inverse* decision rule of $\Phi \to \Psi$.

Let $Dec^*(S)$ denote the set of all inverse decision rules of $Dec(S)$.

It can be shown that $Dec^*(S)$ satisfies 1), 2), 3) and 4), i.e., it is a decision algorithm in S.

If $Dec(S)$ is a decision algorithm then $Dec^*(S)$ will be called an *inverse* decision algorithm of $Dec(S)$.

The inverse decision algorithm gives *reasons* (*explanations*) for decisions pointed out by the decision algorithms.

A decision algorithm is a description of a decision table in the decision language.

Generation of decision algorithms from decision tables is a complex task and we will not discuss this issue here, for it does not lie in the scope of this paper. The interested reader is advised to consult the references.

7 Decision Rules in Information Systems

Decision rules can be also defined, without decision language, referring only to decision tables.

Let $S = (U, C, D)$ be a decision table. Every $x \in U$ determines a sequence $c_1(x), \ldots, c_n(x), d_1(x), \ldots, d_m(x)$ where $\{c_1, \ldots, c_n\} = C$ and $\{d_1, \ldots, d_m\} = D$.

The sequence will be called a *decision rule (induced by x)* in S and denoted by $c_1(x), \ldots, c_n(x) \to d_1(x), \ldots, d_m(x)$ or in short $C \to_x D$.

The number $supp_x(C, D) = |C(x) \cap D(x)|$ will be called a *support* of the decision rule $C \to_x D$ and the number

$$\sigma_x(C, D) = \frac{supp_x(C, D)}{|U|},$$

will be referred to as the *strength* of the decision rule $C \to_x D$. With every decision rule $C \to_x D$ we associate the *certainty factor* of the decision rule, denoted $cer_x(C, D)$ and defined as follows:

$$cer_x(C, D) = \frac{|C(x) \cap D(x)|}{|C(x)|} = \frac{supp_x(C, D)}{|C(x)|} = \\ = \frac{\sigma_x(C, D)}{\pi(C(x))},$$

where $\pi(C(x)) = \frac{|C(x)|}{|U|}$.

The certainty factor may be interpreted as a conditional probability that y belongs to $D(x)$ given y belongs to $C(x)$, symbolically $\pi_x(D|C)$, i.e., $cer_x(C, D) = \pi_x(D|C)$.

If $cer_x(C,D)=1$, then $C \rightarrow_x D$ will be called a *certain decision* rule in S; if $0<cer_x(C,D)<1$ the decision rule will be referred to as an *uncertain decision rule* in S.

The *coverage factor* of the decision rule, denoted $cov_x(C,D)$ is defined as

$$cov_x(C,D)=\frac{|C(x)\cap D(x)|}{|D(x)|}=\frac{supp_x(C,D)}{|D(x)|}= \\ =\frac{\sigma_x(C,D)}{\pi(D(x))},$$

where $\pi(D(x))=\frac{|D(x)|}{|U|}$.

Obviously we have

$$cov_x(C,D)=\pi_x(C|D).$$

If $C \rightarrow_x D$ is a decision rule then $D \rightarrow_x C$ will be called an *inverse decision rule*.

Let us observe that

$$cer_x(C,D)=\mu^C_{D(x)}(x) \text{ and } cov_x(C,D)=\mu^D_{C(x)}(x).$$

That means that the certainty factor expresses the degree of membership of x to the decision class $D(x)$, given C, whereas the coverage factor expresses the degree of membership of x to condition class $C(x)$, given D.

Observe the difference between definitions of decision rules given in section 5 and this section. The previous definition can be regarded as *syntactic* one, whereas the definition given in this section is rather *semantic*.

8 Properties of Decision Rules

Decision rules have important properties which are discussed next.

Let $C \rightarrow_x D$ be a decision rule in S. Then the following properties are valid:

$$\sum_{y\in C(x)} cer_y(C,D)=1 \tag{1}$$

$$\sum_{y\in D(x)} cov_y(C,D)=1 \tag{2}$$

$$\pi(D(x))=\sum_{y\in C(x)} cer_y(C,D)\cdot\pi(C(y))= \\ =\sum_{y\in C(x)}\sigma_y(C,D) \tag{3}$$

$$\pi\left(C\left(x\right)\right)=\sum_{y\in D(x)}cov_{y}\left(C,D\right)\cdot\pi\left(D\left(y\right)\right)= \tag{4}$$
$$=\sum_{y\in D(x)}\sigma_{y}\left(C,D\right)$$

$$cer_{x}\left(C,D\right)=\frac{cov_{x}\left(C,D\right)\cdot\pi\left(D\left(x\right)\right)}{\sum\limits_{y\in D(x)}cov_{y}\left(C,D\right)\cdot\pi\left(D\left(y\right)\right)}= \tag{5}$$
$$=\frac{\sigma_{x}\left(C,D\right)}{\sum\limits_{y\in D(x)}\sigma_{y}\left(C,D\right)}=\frac{\sigma_{x}\left(C,D\right)}{\pi\left(C\left(x\right)\right)}$$

$$cov_{x}\left(C,D\right)=\frac{cer_{x}\left(C,D\right)\cdot\pi\left(C\left(x\right)\right)}{\sum\limits_{y\in C(x)}cer_{y}\left(C,D\right)\cdot\pi\left(C\left(y\right)\right)}= \tag{6}$$
$$=\frac{\sigma_{x}\left(C,D\right)}{\sum\limits_{y\in C(x)}\sigma_{y}\left(C,D\right)}=\frac{\sigma_{x}\left(C,D\right)}{\pi\left(D\left(x\right)\right)}$$

Thus, any decision table, satisfies (1),...,(6). Let us notice that (3) and (4) refer to the well known *total probability theorem*, whereas (5) and (6) refer to *Bayes' theorem*.

Hence in order to compute the certainty and coverage factors of decision rules according to formulas (5) and (6) it is enough to know the strength (support) of all decision rules only.

Let us observe that the above properties are valid also for syntactic decision rules, i.e., any decision algorithm satisfies (1),...,(6). Therefore, in what follows, we will use the concept of the decision table and the decision algorithm equivalently.

9 Decision Tables and Flow Graphs

With every decision table we associate a *flow graph*, i.e., a directed, connected, acyclic graph defined as follows: to every decision rule $C\rightarrow_{x}D$ we assign a *directed branch* x connecting the *input node* $C\left(x\right)$ and the *output node* $D\left(x\right)$. Strength of the decision rule represents a *throughflow* of the corresponding branch. The throughflow of the graph is governed by formulas (1),...,(6). Compare with flow conservation equations in classical network theory [6]

Formulas (1) and (2) are obvious. Formula (3) states that the outflow of the output node amounts to the sum of its inflows, whereas formula (4) says that the sum of outflows of the input node equals to its inflow. Finally, formulas (5) and (6) reveal how throughflow in the flow graph is distributed between its inputs and outputs.

10 Illustrative Examples

Let us illustrate the above ideas by simple examples. These examples intend to show the difference between "classical" Bayesian approach and that proposed by the rough set theory.

Observe that we are not using data to verify prior knowledge, inherently associated with Bayesian data analysis, but the rough set approach shows that any decision table safisties Bayes' theorem and total probability theorem. These properties form the basis of drawing conclusions from data, without referring either to prior or posterior knowledge.

Example 2. This example, which is a modification of example 1 given in section 2, will clearly show the different role of Bayes' theorem in classical statistical inference and that in rough set based data analysis.

Let us consider the data table shown in Table 2.

Table 2. Data table

	T^+	T^-
D	95	5
$\overline{D}$	1998	97902

In Table 2, instead of probabilities, like those given in Table 1, numbers of patients belonging to the corresponding classes are given. Thus we start from the original data (not probabilities) representing outcome of the test.

Now from Table 2 we create a decision table and compute strength of decision rules. The results are shown in Table 3.

Table 3. Decision table

fact	D	T	*support*	*strength*
1	+	+	95	0.00095
2	−	+	1998	0.01998
3	+	−	5	0.00005
4	−	−	97902	0.97902

In Table 3 D is the condition attribute, wheras T is the decision attribute. The decision table is meant to represent a "cause–effect" relation between the disease and result of the test. That is, we expect that the disease causes positive test result and lack of the disease results in negative test result.

The decision algorithm is given below:

1') *if* (*disease, yes*) *then* (*test, positive*)
2') *if* (*disease, no*) *then* (*test, positive*)
3') *if* (*disease, yes*) *then* (*test, negative*)
4') *if* (*disease, no*) *then* (*test, negative*)

The certainty and coverage factors of the decision rules for the above decision algorithm are given is Table 4.

Table 4. Certainty and coverage

rule	*strength*	*certainty*	*coverage*
1	0.00095	0.95	0.04500
2	0.01998	0.02	0.95500
3	0.00005	0.05	0.00005
4	0.97902	0.98	0.99995

The decision algorithm and the certainty factors lead to the following conclusions:

- 95% persons suffering from the disease have positive test result
- 2% healthy persons have positive test result
- 5% persons suffering from the disease have negative test result
- 98% healthy persons have negative test result

That is to say that if a person has the disease most probably the test result will be positive and if a person is healthy the test result will be most probably negative. In other words, in view of the data there is a causal relationship between the disease and the test result.

The inverse decision algorithm is the following:

1) *if* (*test, positive*) *then* (*disease, yes*)
2) *if* (*test, positive*) *then* (*disease, no*)
3) *if* (*test, negative*) *then* (*disease, yes*)
4) *if* (*test, negative*) *then* (*disease, no*)

From the coverage factors we can conclude the following:

- 4.5% persons with positive test result are suffering from the disease
- 95.5% persons with positive test result are not suffering from the disease
- 0.005% persons with negative test result are suffering from the disease
- 99.995% persons with negative test result are not suffering from the disease

That means that if the test result is positive it does not necessarily indicate the disease but negative test result most probably (almost for certain) does indicate lack of the disease.

It is easily seen from Table 4 that (*disease, no*) $\equiv_{0.98,0.02}$ (*test, no*).

That means that the set of all healthy patients and the set of all patients having negative test result is "almost" the same.

That is to say that the negative test result almost exactly identifies healthy patients.

For the remaining rules the accuracy is much smaller and consequently test results are not indicating the presence or absence of the disease.

Example 3. Let us now consider a more complex example, shown in Table 5.

Table 5. Decision table

fact	*disease*	*age*	*sex*	*test*	*support*
1	*yes*	*old*	*man*	+	400
2	*yes*	*middle*	*woman*	+	80
3	*no*	*old*	*man*	−	100
4	*yes*	*old*	*man*	−	40
5	*no*	*young*	*woman*	−	220
6	*yes*	*middle*	*woman*	−	60

Attributes *disease, age* and *sex* are condition attributes, wheras *test* is the decision attribute.

The strength, certainty and coverage factors for decision table are shown in Table 6.

Table 6. Certainty and coverage

fact	*strength*	*certainty*	*coverage*
1	0.44	0.92	0.83
2	0.09	0.56	0.17
3	0.11	1.00	0.23
4	0.04	0.08	0.09
5	0.24	1.00	0.51
6	0.07	0.44	0.15

The flow graph for Table 5 is presented in Fig. 1.

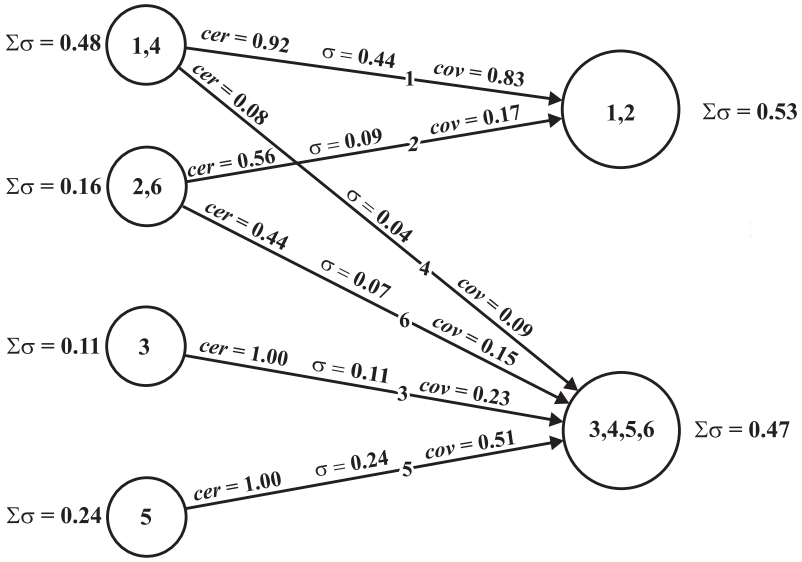

Fig. 1. Flow graph of the decision table

A decision algorithm associated with Table 5 is given below.

1) *if* (*disease, yes*) *and* (*age, old*) *then* (*test*, +)
2) *if* (*disease, yes*) *and* (*age, middle*) *then* (*test*, +)
3) *if* (*disease, no*) *then* (*test*, −)
4) *if* (*disease, yes*) *and* (*age, old*) *then* (*test*, −)
5) *if* (*disease, yes*) *and* (*age, middle*) *then* (*test*, −)

The certainty and coverage factors for the above algorithm are given in Table 7.

Table 7. Certainty and coverage factors

rule	*strength*	*certainty*	*coverage*
1	0.44	0.92	0.83
2	0.09	0.56	0.17
3	0.36	1.00	0.76
4	0.04	0.08	0.09
5	0.07	0.44	0.15

The flow graph for the decision algorithm is presented in Fig. 2.

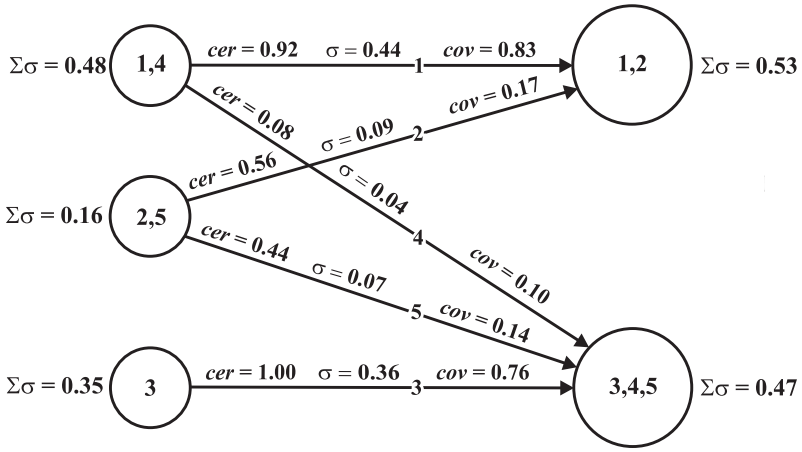

Fig. 2. Flow graph of the decision algorithm

The certainty factors of the decision rules lead to the following conclusions:

- 92% ill and old patients have positive test result
- 56% ill and middle aged patients have positive test result
- all healthy patients have negative test result
- 8% ill and old patients have negative test result
- 44% ill and middle aged patients have negative test result

or in short:

- ill and old patients most probably have positive test result (probability = 0.92)
- ill and middle aged patients most probably have positive test result (probability = 0.56)
- healthy patients have certainly negative test result (probability = 1.00)

From the inverse decision algorithm:

1') *if* (*test*, +) *then* (*disease, yes*) *and* (*age, old*)
2') *if* (*test*, +) *then* (*disease, yes*) *and* (*age, middle*)
3') *if* (*test*, −) *then* (*disease, no*)
4') *if* (*test*, −) *then* (*disease, yes*) *and* (*age, old*)
5') *if* (*test*, −) *then* (*disease, yes*) *and* (*age, middle*)

and the coverage factors we get the following explanation of test results:

- reasons for positive test results are most probably disease and old age (probability = 0.83)

- reason for negative test result is most probably lack of the disease (probability = 0.76)

From the discussed examples it is easily seen the difference between decision tables and decision algorithms. Decision table is a collection of data, whereas a decision algorithm is a linguistic expression, which describes some properties of data in logical (minimal) form.

It follows from Table 6 that there are two interesting approximate equivalences of test results and the disease.

According to rule 1) the disease and old age are approximately equivalent to positive test result ($k = 0.83, \varepsilon = 0.11$), and lack of the disease according to rule 3) is approximately equivalent to negative test result ($k = 0.76, \varepsilon = 0.24$).

It is interesting to examine closely this example but we leave it to the interested reader.

11 Conclusion

From examples 1, 2 and 3 it is easily seen the difference between employing Bayes' theorem in statistical reasoning and the role of Bayes' theorem in rough set based data analysis.

Bayesian inference consists in updating prior probabilities by means of data to posterior probabilities.

In the rough set approach to Bayes' theorem reveals data patterns, which are used next to draw conclusions from data, in form of decision rules.

In other words, classical Bayesian inference is based rather on subjective prior probability, whereas the rough set view on Bayes' theorem refers to objective probability inherently associated with decision tables.

It is also important to notice that in the rough set formulation of Bayes' theorem has a new mathematical form: the conditional probabilitics are expressed in terms of strength of decision rules. This essentially simplifies computations and also gives a new look on Bayesian methodology.

Besides the rough set approach to Bayes' theorem enables us to invert decision rules, i.e. to give reasons for decisions.

Let us also observe that conclusions are valid only for the data set considered. Other data may lead to different conclusions. This is inherent property of inductive reasoning, and reflects the relationship between data sample and the "whole" set of data. This fact is well known not only to philosphers and logicians but also was known to Sherlock Holmes (see Motto).

It seems also important that with every decision table (decision algorithm) a flow graph can be associated, which gives a new tool to decision analysis. The flow graphs considered here are different from those introduced by Ford and Fulkerson and can be formulated in general terms, not associated with decision tables, but this issue has not been considered in this paper.

References

1. Adams, E. W.: The logic of conditionals, an application of probability to deductive Logic. D. Reidel Publishing Company, Dordrecht, Boston (1975)
2. Bayes, T.: An essay toward solving a problem in the doctrine of chances, Phil. Trans. Roy. Soc. **53** (1763) 370–418; Reprint Biometrika **45** (1958) 296–315
3. Bernardo, J. M., Smith, A. F. M.: Baysian theory, Wiley series in probability and mathematical statistics. John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore (1994)
4. Box, G.E.P., Tiao, G.C.: Bayesiaon inference in statistical analysis. John Wiley and Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore (1992)
5. Berthold, M., Hand, D.J.: Intelligent data analysis, an introduction. Springer-Verlag, Berlin , Heidelberg, New York (1999)
6. Ford, L.R., Fulkerson, D. R.: Flows in Networks. Princeton University Press, Princeton, New Yersy (1962)
7. Łukasiewicz, J.: Die logishen Grundlagen der Wahrscheinilchkeitsrechnung. Kraków (1913). In: L. Borkowski (ed.), Jan Łukasiewicz – Selected Works, North Holland Publishing Company, Amsterdam, London, Polish Scientific Publishers, Warsaw (1970)
8. Pawlak, Z.: Rough sets – theoretical aspect of reasoning about data, Kluwer Academic Publishers, Boston, Dordrech, London (1991)
9. Pawlak, Z.: New Look on Bayes' Theorem – the Rough Set Outlook. Proceedings of the International Workshop on Rough Set Theory and Granular Computing (RSTGC-2001), S. Hirano, M. Inuiguichi and S. Tsumoto (eds.), Bull. of International Rough Set Society, Vol. 5. No. 1/2, Matsue, Shimane, Japan, May 20-22, (2001) 1–8
10. Pawlak, Z., Skowron, A.: Rough membership functions. Advances in the Dempster-Shafer Theory of Evidence, R, Yager, M. Fedrizzi, J. Kacprzyk (eds.), John Wiley & Sons, Inc. New York (1994) 251–271
11. Skowron, A.: Rough Sets in KDD (plenary talk); 16-th World Computer Congress (IFFIP'2000), Beijing, August 19-25, 2000, In:Zhongzhi Shi, Boi Faltings, Mark Musem (eds.) Proceedings of the Conference on Intelligent Information Processing (IIP2000), Publishing Hous of Electronic Industry, Beijing (2000) 1–17
12. Tsumoto, S., Tanaka, H.: Discovery of functional components of proteins based on PRIMEROSE and domain knowledge hierarchy. Proceedings of the Workshop on Rough Sets and Soft Computing (RSSC-94) (1994): Lin, T.Y., and Wildberger, A.M.(eds.) Soft Computing (1995) 280–285