The Rough Set View on Bayes' Theorem

Zdzisław Pawlak

University of Information Technology and Management
ul. Newelska 6, 01–447 Warsaw, Poland
zpw@ii.pw.edu.pl

MOTTO:
"It is a capital mistake to theorise before one has data"
Sherlock Holms
In: A Scandal in Bohemia

Abstract. Rough set theory offers new perspective on Bayes' theorem. The look on Bayes' theorem offered by rough set theory reveals that any data set (decision table) satisfies total probability theorem and Bayes' theorem. These properties can be used directly to draw conclusions from objective data without referring to subjective prior knowledge and its revision if new evidence is available.
Thus the rough set view on Bayes' theorem is rather objective in contrast to subjective "classical" interpretation of the theorem .

1 Introduction

In his paper [2] Bayes considered 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."

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

Currently Bayes' theorem is the basic of statistical interference.

"The result of the Bayesian data analysis process is the posterior distribution that represents a revision of the prior distribution on the light of the evidence provided by the data" [5].

Bayes' based inference methodology rised many controversy and criticism. For example,

"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].

"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].

Rough set theory offers new insight into Bayes' theorem. The look on Bayes' theorem offered by rough set theory is completely different to that used in the

N.R. Pal and M. Sugeno (Eds.): AFSS 2002, LNAI 2275, pp. 106–116, 2002.

Bayesian data analysis philosophy. It 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. It states that any data set (decision table) satisfies total probability theorem and Bayes' theorem. This property can be used directly to draw conclusions from the data without referring to prior knowledge and its revision if new evidence is available. Thus in the presented approach the only source of knowledge is the data and there is no need to assume that there is any prior knowledge besides the data.

Moreover, the rough set approach to Bayes' theorem shows close relationship between logic of implications and probability, which was first observed by Łukasiewicz [6] 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. This is a very important feature of utmost importance to data mining and decision analysis, for it extends the class of problem which can be considered in these domains.

Besides, we propose a new form of Bayes' theorem where basic role plays 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 method of inference and also simplifies essentially computations.

It is also interesting to note a relationship between Bayes' theorem and flow graphs.

Let us also observe that the rough set view on Bayes' theorem is rather objective in contrast to subjective "classical" interpretation.

2 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 [7, 10].

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 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.

3 Rough Membership

Rough sets can be also defined employing instead of approximations rough membership function [9], 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 $X \subseteq U$ and $B \subseteq A$ 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.

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 $x \in U$

Compare these properties to those of fuzzy membership. Obviously rough membership is a generalization of fuzzy membership.

The rough membership function, can be used to define approximations and the boundary region of a set, as shown below:

$$B_*(X) = \{x \in U : \mu_X^B(x) = 1\},$$

$$B^*(X) = \{x \in U : \mu_X^B(x) > 0\},$$

$$BN_B(X) = \{x \in U : 0 < \mu_X^B(x) < 1\}.$$

4 Information Systems and Decision Rules

Every decision table describes decisions determined, 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.

Let us describe decision rules more exactly.

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 $\pi_x(D|C)$.

If $cer_x(C, D) = 1$, then $C \to_x D$ 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.

Besides, we will also use a *coverage factor* of the decision rule, denoted $cov_x(C,D)$ 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 $\pi(D(x)) = \frac{|D(x)|}{|U|}$.

Similarly

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

If $C \to_x D$ is a decision rule then $D \to_x C$ will be called an *inverse decision rule*. The inverse decision rules can be used to give *explanations* (*reasons*) for decisions.

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.

Decision rules are often represented in a form of "*if ... then ...*" implications. Thus any decision table can be transformed in a set of "*if ... then ...*" rules, called a *decision algorithm*.

Generation of minimal decision algorithms from decision tables is a complex task and we will not discuss this issue here. The interested reader is advised to consult the references.

5 Probabilistic Properties of Decision Tables

Decision tables have important probabilistic properties which are discussed next.

Let $C \to_x D$ be a decision rule in S and let $\Gamma = C(x)$ and let $\Delta = D(x)$. Then the following properties are valid:

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

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

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

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

$$cer_x\left(C, D\right) = \frac{cov_x\left(C, D\right) \cdot \pi\left(D\left(x\right)\right)}{\sum\limits_{y \in \Delta} cov_y\left(C, D\right) \cdot \pi\left(D\left(y\right)\right)} = \tag{5}$$
$$= \frac{\sigma_x\left(C, D\right)}{\pi\left(C\left(x\right)\right)}$$

$$cov_x\left(C, D\right) = \frac{cer_x\left(C, D\right) \cdot \pi\left(C\left(x\right)\right)}{\sum\limits_{y \in \Gamma} cer_y\left(C, D\right) \cdot \pi\left(C\left(y\right)\right)} = \tag{6}$$
$$= \frac{\sigma_x\left(C, D\right)}{\pi\left(D\left(x\right)\right)}$$

That is, any decision table, satisfies (1),...,(6). Observe that (3) and (4) refer to the well known *total probability theorem*, whereas (5) and (6) refer to *Bayes' theorem*.

Thus in order to compute the certainty and coverage factors of decision rules according to formula (5) and (6) it is enough to know the strength (support) of all decision rules only. The strength of decision rules can be computed from data or can be a subjective assessment.

6 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).

Formulas (1) and (2) say that an outflow of an input node or an output node is equal to their inflows. 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.

7 Illustrative Examples

Now we will illustrate the ideas considered in the previous sections by simple tutorial examples. These examples intend to show clearly the difference between "classical" Bayesian approach and that proposed by the rough set philosophy.

Example 1. This example 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 1.

Table 1. Data table

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

In Table 1 the number of patients belonging to the corresponding classes is given. Thus we start from the original data (not probabilities) representing outcome of the test.

Now from Table 1 we create a decision table and compute strength of decision rules. The results are shown in Table 2.

Table 2. 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 2 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 3.

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

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

Table 3. 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

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 results are suffering from the disease
- 99.995% persons with negative test results 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 results most probably (almost for certain) does indicate lack of the disease.

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 2. Let us now consider a little more sophisticated example, shown in Table 4.

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 5.

Below a decision algorithm associated with Table 5 is presented.

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*, −)

Table 4. 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

Table 5. 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

5) *if* (*disease, yes*) *and* (*age, middle*) *then* (*test*, −)

The flow graph for the decision algorithm is presented in Fig. 1.

The certainty and coverage factors for the above algorithm are given in Table 6

Table 6. 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 certainty factors of the decision rules lead the following conclusions:

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

In other words:

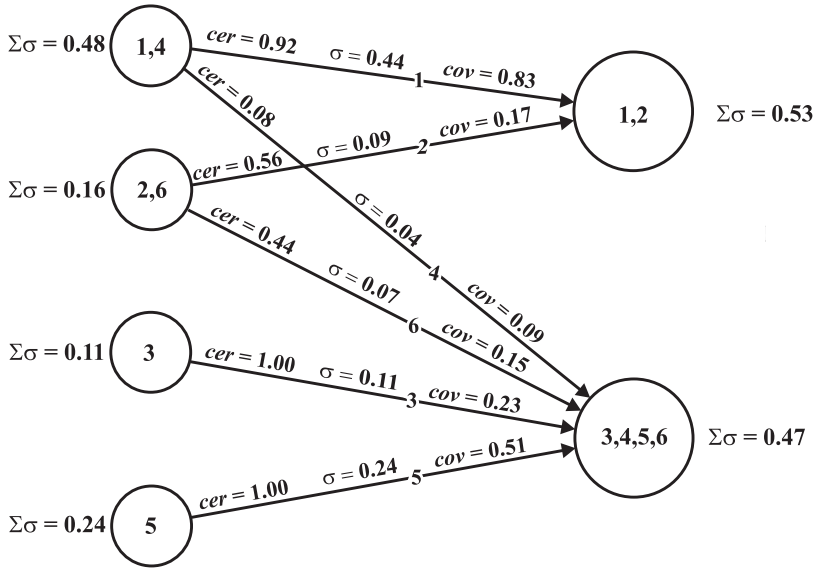

Fig. 1. Flow graf

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

Now let us examine the inverse decision algorithm, which is given below:

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*)

Employing the inverse decision algorithm and the coverage factor 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) □

If is clearly seen from examples 1 and 2 the difference between Bayesian data analysis and the rough set approach. In the Bayesian inference the data is used to update prior knowledge (probability) into a posterior probability, whereas rough sets are used to understand what the data are telling us.

8 Conclusion

Bayesian inference consists in updating prior probabilities by means of data to posterior probabilities, which is rather subjective.

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

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.: Bayesian 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. 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)
7. Pawlak, Z.: Rough Sets – Theoretical Aspect of Reasoning about Data. Kluwer Academic Publishers, Boston Dordrech, London (1991)
8. Pawlak, Z.: New look on Bayes' theorem – the rough set outlook. Proceeding of International Workshop on Rough Set Theory and Granular Computing (RSTGC-2001), Matsue, Shimane, Japan, May 20-22, S. Hirano, M. Inuiguchi and S. Tsumoto (eds.), Bull. of Int. Rough Set Society vol. **5** no. **1/2** 2001 1–8
9. Z. Pawlak, A. Skowron, 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