27. Bayes' Theorem Revised – The Rough Set View

Zdzisław Pawlak

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

Rough set theory offers new insight into Bayes' theorem. The look on Bayes' theorem offered by rough set theory is completely different from that used in the 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 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. We simply look what the data are telling us. Consequently we do not refer to any prior knowledge which is updated after receiving some data.

27.1 Introduction

This paper is an abbreviation of [27.8]

Bayes' theorem is the essence of statistical inference.

"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" [27.5].

"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" [27.4].

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 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 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. We simply look what the data are telling us. Consequently

T. Terano et al. (Eds.): JSAI 2001 Workshops, LNAI 2253, pp. 240–250, 2001.

we do not refer to any prior knowledge which is updated after receiving some data.

Moreover, the rough set approach to Bayes' theorem shows close relationship between logic of implications and probability, which was first observed by Łukasiewicz [27.6] and also independly studied by Adams [27.1] 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 an subjective assessment. This formulation gives new look on Bayesian method of inference and also essentially simplifies computations.

27.2 Bayes' Theorem

In this section we recall basic ideas of Bayesian inference philosophy, after recent books on Bayes' theory citeber:smi,box:tia,bert:han.

In his paper [27.2] 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."

"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" [27.3].

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

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

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*" [27.3].

"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" [27.4].

Let us illustrate the above by a simple example taken from [27.5].

Example 27.2.1. "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 27.1. 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". □

27.3 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 [27.7, 27.10].

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 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.

27.4 Rough Membership

Rough sets can be also defined employing instead of approximations rough membership function [27.9], 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 $B \subseteq A$.

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

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

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

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

27.5 Information Systems and Decision Rules

Every decision table describes decisions (actions, results etc.) 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.

In what follows we will 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 $C \to_x D$.

Decision rules are often presented as logical implications in the form "*if...then...*".

A set of decision rules corresponding to a decision table will be called a *decision algorithm.*

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 $C \to_x D$, where $|X|$ denotes the cardinality of X. 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:

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

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

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

If $cer_x\left(C,D\right)=1$, then $C\rightarrow_x D$ will be called a *certain decision* rule in S; if $0<cer_x\left(C,D\right)<1$ 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 $cov_x\left(C,D\right)$ defined as

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

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

Similarly

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

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

Let us observe that

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

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

27.6 Probabilistic Properties of Decision Tables

Decision tables have important probabilistic properties which are discussed next.

Let $C\rightarrow_x D$ be a decision rule in S and let $\varGamma=C\left(x\right)$ and let $\varDelta=D\left(x\right)$. Then the following properties are valid:

$$\sum_{y\in\varGamma}cer_y\left(C,D\right)=1 \tag{27.1}$$

$$\sum_{y \in \varDelta} cov_y\left(C, D\right) = 1 \tag{27.2}$$

$$\begin{aligned} \pi\left(D\left(x\right)\right) &= \sum_{y \in \varGamma} cer_y\left(C, D\right) \cdot \pi\left(C\left(y\right)\right) = \\ &= \sum_{y \in \varGamma} \sigma_y\left(C, D\right) \end{aligned} \tag{27.3}$$

$$\begin{aligned} \pi\left(C\left(x\right)\right) &= \sum_{y \in \varDelta} cov_y\left(C, D\right) \cdot \pi\left(D\left(y\right)\right) = \\ &= \sum_{y \in \varDelta} \sigma_y\left(C, D\right) \end{aligned} \tag{27.4}$$

$$\begin{aligned} cer_x\left(C, D\right) &= \frac{cov_x\left(C, D\right) \cdot \pi\left(D\left(x\right)\right)}{\sum_{y \in \varDelta} cov_y\left(C, D\right) \cdot \pi\left(D\left(y\right)\right)} = \\ &= \frac{\sigma_x\left(C, D\right)}{\pi\left(C\left(x\right)\right)} \end{aligned} \tag{27.5}$$

$$\begin{aligned} cov_x\left(C, D\right) &= \frac{cer_x\left(C, D\right) \cdot \pi\left(C\left(x\right)\right)}{\sum_{y \in \varGamma} cer_y\left(C, D\right) \cdot \pi\left(C\left(y\right)\right)} = \\ &= \frac{\sigma_x\left(C, D\right)}{\pi\left(D\left(x\right)\right)} \end{aligned} \tag{27.6}$$

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 formulas (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.

Let us observe that the above properties are valid also for syntactic decision rules, i.e., any decision algorythm satisfies (1),...,(6).

Thus, in what follows, we will use the concept of the decision table and the decision algorithm equivalently.

27.7 Decision Tables and Flow Graphs

With every decision table we associate a *flow graph*, i.e., a directed 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.

27.8 Comparison of Bayesian and Rough Set Approach

Now we will illustrate the ideas considered in the previous sections by means of the example considered in section 2. These examples intend to show clearly the difference between "classical" Bayesian approach and that proposed by the rough set philosophy.

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 27.8.1. 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 27.2. 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) represanting 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.

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.

Table 27.3. Decision table

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

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 27.4. 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 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

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.

It is easily seen from Table 4 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. □

It 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.

27.9 Conclusion

From examples 1 and 2 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 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.

Acknowledgments. The author wishes to express his gratitude to Professor Andrzej Skowron for many critical remarks.

References

27.1 Adams, E. W.: The logic of conditionals, an application of probability to deductive Logic. D. Reidel Publishing Company, Dordrecht, Boston (1975)

27.2 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

27.3 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)

27.4 Box, G.E.P., Tiao, G.C.: Bayesiaon inference in statistical analysis. John Wiley and Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore (1992)

27.5 Berthold, M., Hand, D.J.: Intelligent data analysis, an introduction. Springer-Verlag, Berlin , Heidelberg, New York (1999)

27.6 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)

27.7 Pawlak, Z.: Rough sets – theoretical aspect of reasoning about data, Kluwer Academic Publishers, Boston, Dordrech, London (1991)

27.8 Pawlak, Z.: New look on Bayes' theorem – the rough set outlook. In: S. Hirano, M. Inuiguchi, S. Tsumoto (eds.), Proceedings of the International Workshop on Rough Set Theory and Granular Computing (RSTGC-2001), Ball of the International Set Society, Vol. 5 No.1/2, Matsue, Shimane, Japan, May 20-22 (2001) 1–8

27.9 Pawlak, Z.: Rough sets and decision algorithms. Springer-Verlag, Berlin, Heidelberg, New York (to appear)

27.10 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. ew York (1994) 251–271

27.11 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

27.12 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