B. Bouchon-Meunier, L. Foulloy and R.R. Yager (Editors)
Intelligent Systems for Information Processing:
From Representation to Applications

Rough Sets, Bayes' Theorem and Flow Graphs

Zdzisław Pawlak
Institute for Theoretical and Applied Informatics
Polish Academy of Sciences
ul. Bałtycka 5, 44 100 Gliwice, Poland
e-mail: zpw@ii.pw.edu.pl

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

Abstract

Rough set theory is a new approach to vagueness and uncertainty. The theory of rough sets has an overlap with many other theories. Specially interesting is the relationship to fuzzy set theory and the theory of evidence. Recently, it turned out that the theory has very interesting connections with 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. Besides, it is revealed that Bayes' theorem can be interpreted as a flow *conservation equation* in a flow graph. However the flow graphs considered here are different from those introduced by Ford and Fulkerson. This property gives new perspective for applications of Bayes' theorem. Thus the paper brings two new interpretation of Bayes' theorem, without referring to its classical probabilistic interpretation: as properties of data tables and properties of flow graphs.

Keywords: roughs sets, Bayes' theorem, decision rules, flow graphs.

1 Introduction

Rough set theory is a new approach to vagueness and uncertainty. Foundation of rough sets can be found in [9]. The theory has found many applications, in particular in data analysis and data mining, offering new look and tools for these domains.

The theory of rough sets has an overlap with many other theories. Specially interesting is the relationship to fuzzy set theory and the theory of evidence. Recently, it turned out that the theory has very interesting connections with Bayes' theorem. This link gives a new look on Bayes' theorem which is significant not only from philosophical point of view but also offers new methods of data analysis.

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 [6, 7, 8].
It is also interesting that Bayes' theorem can be interpreted as a *flow conservation equation* in a flow graph. However the flow graphs considered here are different from those introduced by Ford and Fulkerson [4].

2 Rough Set Theory - Basic Concepts

In this section we define basic concepts of rough set theory: information system and approximation of sets.
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., $BN_B(X) = \varnothing$, then *X* is *crisp* (*exact*) with respect to *B*; in the opposite case, i.e., if $BN_B(X) \neq \varnothing$, *X* is referred to as *rough* (*inexact*) with respect to *B*.
Rough sets can be also defined employing instead of approximations rough membership function, which is defined as follows:

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

and

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

where $X \subseteq U$, $B \subseteq A$ and *card*(*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 [9]:

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. $\mu_{X \cup Y}(x) \geq max\left(\mu_X^B(x), \mu_Y^B(x)\right)$ for any $x \in U$
6. $\mu_{X \cap Y}(x) \leq min\left(\mu_X^B(x), \mu_Y^B(x)\right)$ 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:

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

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

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

3 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 $c_1(x),\ldots,c_n(x)\ d_1(x),\ldots,d_m(x)$, in short $C \rightarrow_x D$.

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

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

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

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

where $\sigma_x(C) = \dfrac{card(C(x))}{card(U)}$.

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

Besides, we will also use a *coverage factor* of the decision rule, denoted $cov_x(C, D)$ and defined as

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

where $\sigma_x(D) = \dfrac{card(D(x))}{card(U)}$.

Similarly

$$cov_x(C,D) = \sigma_x(C \mid D).$$

The certainty and coverage factors have been for a long time used in machine learning and data mining, [11, 12] but in fact they have been first introduced in 1913 by Jan Łukasiewicz, in connection with his study of logic and probability [5].

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

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

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 the 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 rather difficult. Many methods for solving this problem have been proposed but we will not discuss this issue in this paper.

4 Properties of Decision Rules

Decision rules have important properties which are discussed below.
Let $C \rightarrow_x D$ be a decision rule in S. Then the following properties are valid:

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

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

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

$$\sigma_x(C) = \sum_{y \in D(x)} cov_y(C,D) \cdot \sigma_x(D) = \sum_{y \in D(x)} \sigma_y(C,D) \tag{4}$$

$$cer_x(C,D) = \frac{cov_x(C,D) \cdot \sigma_x(D)}{\sum_{y \in D(x)} cov_y(C,D) \cdot \sigma_x(D)} = \frac{\sigma_x(C,D)}{\sum_{y \in D(x)} \sigma_y(C,D)} = \frac{\sigma_x(C,D)}{\sigma_x(C)} \tag{5}$$

$$cov_x(C,D) = \frac{cer_x(C,D) \cdot \sigma_x(C)}{\sum_{y \in C(x)} cer_y(C,D) \cdot \sigma_x(C)} = \frac{\sigma_x(C,D)}{\sum_{y \in C(x)} \sigma_y(C,D)} = \frac{\sigma_x(C,D)}{\sigma_x(D)} \tag{6}$$

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.

5 Dependences in decision Tables

Next important issue in decision table analysis is the dependency of attributes, particularly dependency of decision attributes on condition attributes [10].

Intuitively speaking the set of decision attributes depends on the set of condition attributes if values of decision attributes are totally (or partially) determined by values of condition attributes.
In other words, this kind of dependency describes which decisions specified in a decision table are to by obeyed if some conditions are satisfied.
In this paper we will introduce another kind of dependences in decision tables, based on some ideas of statistics.
Let $S = (U,C,D)$ be a decision table and let $x \in U$. We say that decisions $d_1(x),\ldots,d_m(x)$ are *independent* on conditions $c_1(x),\ldots,c_n(x)$, where $C = \{c_1,\ldots,c_n\}$, $D = \{d_1,\ldots,d_m\}$, if

$$\sigma_x(C,D) = \sigma_x(C)\sigma_x(D).$$

In other words conditions and decisions in a decision rule $C \rightarrow_x D$ are independent if

$$\frac{\sigma_x(C,D)}{\sigma_x(C)} = cer_x(C,D) = \sigma_x(D)$$

or

$$\frac{\sigma_x(C,D)}{\sigma_x(D)} = cov(C,D) = \sigma_x(C).$$

If

$$cer_x(C,D) > \sigma_x(D)$$

or

$$cov_x(C,D) > \sigma_x(C).$$

We say that *D depends positively on C* in a decision rule $C \rightarrow_x D$. Similarly if

$$cer_x(C,D) < \sigma_x(D)$$

or

$$cov_x(C,D) < \sigma_x(C).$$

We say that *D depends negatively* on *C* in a decision rule $C \rightarrow_x D$.
Obviously, the relations of independence and positive and negative dependence are symmetric ones.
Again borrowing from statistics the idea of a correlation coefficient we can determine the degree of dependency numerically, defining the *correlation factor*, defined as follows:

$$\eta_x(C,D) = \frac{cer_x(C,D) - \sigma_x(C)}{cer_x(C,D) + \sigma_x(C)} = \frac{cov_x(C,D) - \sigma(D)}{cov_x(C,D) + \sigma_x(D)}.$$

Obviously, $0 \le \eta_x(C,D) \le 1$ and if $\eta_x(C,D) = 0$ then C and D are independent, if $\eta_x(C,D) < 0$ then C and D are negatively dependent if $\eta_x(C,D) > 0$, then C and D are positively dependent.

6 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* $C(x)$ and the *output node* $D(x)$. Strength of the decision rule

represents a *throughflow* of the corresponding branch. The throughflow of the graph is governed by formulas (1),...,(6), and can be considered as a *flow conservation equation* similar to that introduced by Ford and Fulkerson [4]. However, let us observe that flow graphs presented in this paper are different from flow networks of Ford and Fulkerson.
Formulas (1) and (2) say that the 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.
It is obvious that the idea of flow graph can be also formulated more generally, independently of decision tables, but we will not consider this issue here.

7 An Example

Let us now illustrate the above ideas by means of a simple example shown in Table 1.

Table 1: Decision table

FACT	DIS.	AGE	SEX	TEST	SUPP.
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, whereas **TEST** is the decision attribute.
Below a decision algorithm associated with Table 1 is presented.

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

The certainty and coverage factors for the above algorithm are given in Table 2.

Table 2: Certainty and coverage factors

RULE	STRENGTH	CER.	COV.
1	0.44	0.90	0.83
2	0.09	0.56	0.17
3	0.35	1.00	0.77
4	0.04	0.08	0.09
5	0.07	0.44	0.15

Remark. Due to the round-off errors in computations the properties (1)...(6) may not always be satisfied in the table.

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

- 90% 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

In other words:

- ill and old patients most probably have positive test result (probability = 0.90)
- middle aged patients most probably have positive test result (probability = 0.56)
- healthy patients have certainly negative test result (probability = 1.00)

The inverse decision algorithm is given below:

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

Employing the inverse decision algorithm and the coverage factors we get the following explanation of test result:

- 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 = 0.77)

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

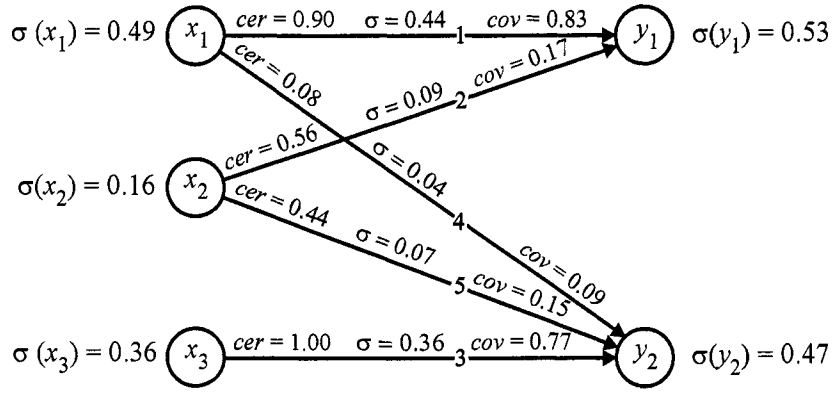

Figure 1: Flow graph

In the flow graph we have $x_1 = \{1, 4\}, x_2 = \{2, 6\}, x_3 = \{3, 5\}, y_1 = \{1, 2\}$ and $y_2 = \{3, 4, 5, 6\}$.

That means that we have in our data base the following groups of patients:

440 – ill and old (x_1)
140 – middle aged (x_2)
320 – healthy (x_3)
and
480 – with positive test result (y_1)
420 – with negative test result (y_2)

Each input node of the flow graph represents a condition of corresponding decision rule, whereas each output node reveals decisions of the rules. The associated numbers can be understood as probabilities (frequencies) of conditions and decisions respectively. Branches of the graph are labeled by strength of associated decision rules.

The flow graph of a decision algorithm shows how probabilities of decisions and conditions are related.

Each node of the graph satisfies equations (1)...(6). Observe, that in order to compute all the conditional and total probabilities it is enough to know the strength of the decision rules only, which makes the computations very easy, and gives also clear insight into the structure of the decision algorithms.

Let us notice that, for example, old age and illness are positively correlated with positive test result, whereas old age and illness are negatively correlated with negative test result.

The corresponding correlation coefficients are 0.26 and – 0.71 respectively.

8 Conclusion

It is clearly seen from the above considerations the difference between Bayesian data analysis and the rough set approach. In the Bayesian inference the data is used to update prior probability (knowledge) into a posterior probability, whereas rough set based Bayesian inference is used to reason directly from data.

The relationship of rough set theory with Bayes' theorem and flow graphs gives new look on Bayesian inference and leads to efficient algorithms for data analysis.

References

[1] J. M. Bernardo, A. F. M. Smith (1994). Bayesian Theory, Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore.

[2] G.E.P. Box, G.C. Tiao (1992). Bayesian Inference in Statistical Analysis. John Wiley and Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore.

[3] M. Berthold, D.J Hand (1999). Intelligent Data Analysis, an Introduction. Springer-Verlag, Berlin, Heidelberg, New York.

[4] L.R. Ford, D. R. Fulkerson (1962). Flows in Networks. Princeton University Press, Princeton, New Jersey.

[5] J. Łukasiewicz (1970). 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.

[6] Z. Pawlak (2001). New Look on Bayes' Theorem - the Rough Set Outlook. In *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, pages 1-8.

[7] Z. Pawlak (2002). In Pursuit of Patterns in Data Reasoning from Data – The Rough Set Way. In: J.J. Alpigini *et al.* (eds.), *Lecture Notes in Artificial Intelligence* 2475, pages 1-9.

[8] Z. Pawlak (2002). Rough Sets, Decision Algorithms and Bayes' Theorem. In *European Journal of Operational Research* 136, pages 181-189.

[9] Z. Pawlak, A. Skowron (1994). Rough Membership Functions. Advances in the Dempster-Shafer Theory of Evidence. R, Yager, M. Fedrizzi, J. Kacprzyk (eds.), John Wiley & Sons, Inc., New York, pages 251-271.

[10] L. Polkowski (2002). Rough Sets – Mathematical Foundations. Physical-Verlag, Springer Verlag Company.

[11] S. Tsumoto, H. Tanaka (1995). Discovery of Functional Components of Proteins Based on PRIMEROSE and Domain Knowledge Hierarchy. In *Proceedings of the Workshop on Rough Sets and Soft Computing (RSSC-94)*, 1994: Lin, T.Y., and Wildberger, A.M. (eds.), *Soft Computing*, SCS 280-285.

[12] S.K.M.Wong, W. Ziarko(1986). Algorithm for Inductive Learning. In *Bull. Polish Academy of Sciences* 34, 5-6, pages 271-276.