A Rough Set View on Bayes' Theorem

Zdzisław Pawlak*
University of Information Technology and Management, ul. Newelska 6,
01 447 Warsaw, Poland

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 the 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. © 2003 Wiley Periodicals, Inc.

MOTTO:
"It is a capital mistake to theorise before one has data"
Sherlock Holmes
(*A Scandal in Bohemia*)

1. INTRODUCTION

This article is a continuation of some ideas presented in Ref. 1, which were inspired by the research of Łukasiewicz on the relationship between logic and probability.[2] He first pointed out connections between implication and Bayes' theorem. His observation is the starting point of the approach presented in this study. Before we start our considerations, let us briefly recall some facts and opinions about Bayes' theorem.

In his article[3] 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) form".[4]

Bayes' theorem has the following form

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

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

© 2003 Wiley Periodicals, Inc. Published online in Wiley InterScience
(www.interscience.wiley.com). • DOI 10.1002/int.10099

where $P(H)$ is called the prior probability of a hypothesis H and expresses our knowledge about H before having data D; $P(H|D)$—called the posterior probability of H given D—tells us what is known about H after obtaining data.

Currently, Bayes' theorem is the basis 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 raised much controversy and criticism, e.g., "Opinion as to the values of Bayes' theorem as a basis for statistical inference has swung between acceptance and rejection since its publication in 1763."[6]

"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".[4]

Rough set theory offers new insight into Bayes' theorem. The look on Bayes' theorem presented here 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.

Moreover, the proposed approach to Bayes' theorem shows a close relationship between logic of implications and probability, which was first observed by Łukasiewicz[2] and also independently studied by Adams[7] 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, because it extends the class of problem, which can be considered in these domains.

Besides, we propose a new form of Bayes' theorem in which the 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 also can be a subjective assessment. This formulation gives a new look on Bayesian method of inference and also simplifies essentially computations.[8]

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 Ref. 1.

An information system is a data table in which its columns are labeled by attributes, rows are labeled by objects of interest, and entries of the table are attribute values.

Formally, by an information system we will understand a pair $S = (U, A)$, where U and A are finite, nonempty sets called the universe and the set of attributes, respectively. With every attribute $a \in A$ we associate a set V_a of its values, called the domain of a. Any subset B of A determines a binary relation $I(B)$ on U, which will be called an indiscernibility relation and is defined as follows: $(x, y) \in I(B)$ if and only if $a(x) = a(y)$ for every $a \in A$, where $a(x)$ denotes the value of attribute a for element x. Obviously, $I(B)$ is an equivalence relation. The family of all equivalence classes of $I(B)$, i.e., 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 that must be taken when some conditions are satisfied. In other words, each row of the of the decision table specifies a decision rule that determines decisions in terms of conditions.

Observe that elements of the universe, in the case of decision tables, are 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 are 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 also can be defined using rough membership function instead of approximations,[9] which is defined as follows:

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

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) $\mu_{X \cup Y}^B(x) \geqslant \max(\mu_X^B(x), \mu_Y^B(x))$ for any $x \in U$
(6) $\mu_{X \cap Y}^B(x) \leqslant \min(\mu_X^B(x), \mu_Y^B(x))$ for any $x \in U$

Compare these properties with 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 that 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 $c_1(x), \ldots, c_n(x) \rightarrow d_1(x), \ldots, d_m(x)$ or, in short, $C \rightarrow_x D$.

The number $\mathrm{supp}_x(C, D) = |C(x) \cap D(x)|$ will be called a support of the decision rule $C \rightarrow_x D$ and the number

$$\sigma_x(C, D) = \frac{\mathrm{supp}_x(C, D)}{|U|}$$

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 $\mathrm{cer}_x(C, D)$ and defined as follows:

$$\mathrm{cer}_x(C, D) = \frac{|C(x) \cap D(x)|}{|C(x)|} = \frac{\mathrm{supp}_x(C, D)}{|C(x)|} = \frac{\sigma_x(C, D)}{\pi(C(x))}$$

where $\pi(C(x)) = [|C(x)|/|U|]$.

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 $\mathrm{cer}_x(C, D) = 1$, then $C \rightarrow_x D$ will be called a certain decision rule in S; if $0 < \mathrm{cer}_x(C, D) < 1$, the decision rule will be referred to as an uncertain decision rule in S. Furthermore, we will also use a coverage factor[10] of the decision rule, denoted $\mathrm{cov}_x(C, D)$, defined as

$$\mathrm{cov}_x(C, D) = \frac{|C(x) \cap D(x)|}{|D(x)|} = \frac{\mathrm{supp}_x(C, D)}{|D(x)|}$$

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

where $\pi(D(x)) = [|D(x)|/|U|]$.

Similarly,

$$\mathrm{cov}_x(C, D) = \pi_x(C|D)$$

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

Let us observe that

$$\mathrm{cer}_x(C, D) = \mu^{C}_{D(x)}(x) \quad \text{and} \quad \mathrm{cov}_x(C, D) = \mu^{D}_{C(x)}(x)$$

Moreover, it is worthwhile to mention that the certainty factor of a decision rule $C \rightarrow_x D$ is, in fact, the coverage factor of the inverse decision rule $D \rightarrow_x C$.

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 often are 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, e.g., Ref. 11.

5. 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 $\Gamma = C(x)$ and $\Delta = D(x)$. Then, the following properties are valid:

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

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

$$\begin{aligned} \pi(D(x)) &= \sum_{y \in \Gamma} \operatorname{cer}_y(C, D) \cdot \pi(C(y)) \\ &= \sum_{y \in \Gamma} \sigma_y(C, D) \end{aligned} \tag{3}$$

$$\begin{aligned} \pi(C(x)) &= \sum_{y \in \Delta} \operatorname{cov}_y(C, D) \cdot \pi(D(y)) \\ &= \sum_{y \in \Delta} \sigma_y(C, D) \end{aligned} \tag{4}$$

$$\begin{aligned} \operatorname{cer}_x(C, D) &= \frac{\operatorname{cov}_x(C, D) \cdot \pi(D(x))}{\sum_{y \in \Delta} \operatorname{cov}_y(C, D) \cdot \pi(D(y))} \\ &= \frac{\sigma_x(C, D)}{\pi(C(x))} \end{aligned} \tag{5}$$

$$\begin{aligned} \operatorname{cov}_x(C, D) &= \frac{\operatorname{cer}_x(C, D) \cdot \pi(C(x))}{\sum_{y \in \Gamma} \operatorname{cer}_y(C, D) \cdot \pi(C(y))} \\ &= \frac{\sigma_x(C, D)}{\pi(D(x))} \end{aligned} \tag{6}$$

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

Thus, to compute the certainty and coverage factors of decision rules according to Equations 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 $C(x)$ and the output node $D(x)$. With every branch x of the flow graph, we associate the number $\gamma(x) = \sigma_x(C, D)$, called throughflow of x. The throughflow of the graph is governed by Equations 1–6.

Table I. Data table.

	T^+	T^-
D	95	5
$\bar{D}$	1998	97,902

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

Observe that Equations 1–6 can be regarded as flow conservation equations similar to those introduced by Ford and Fulkerson[12] in flow networks. However, flow graphs introduced in this study differ essentially from flow networks and refer to other aspects of flow than those considered in Ref. 12.

In addition, let us mention that the concept of the flow graph can be formulated in more general terms, independently from decision tables and it can find quite new applications.

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 show clearly 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 I. In Table I 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 I we create a decision table and compute strength of decision rules. The results are shown in Table II. In Table II, D is the condition attribute, whereas T is the decision attribute. The decision table is meant to represent a "cause-effect" relation between the disease and the result of the test. That is, we expect that the disease causes a positive test result and lack of the disease results in negative test result.

The decision algorithm is given below:

Table II. Decision table.

Fact	D	T	Support	Strength
1	+	+	95	0.00095
2	−	+	1998	0.01998
3	+	−	5	0.00005
4	−	−	97,902	0.97902

Table III. 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

(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 foregoing decision algorithm are given in Table III. The decision algorithm and the certainty factors lead to the following conclusions:

- Ninety-five percent of persons suffering from the disease have positive test results
- Two percent of healthy persons have positive test results
- Five percent of persons suffering from the disease have negative test result
- Ninety-eight percent of 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 most probably will be 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:

- That 4.5% of persons with positive test result are suffering from the disease
- That 95.5% of persons with positive test result are not suffering from the disease

Table IV. 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 V. 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	0.24
4	0.04	0.08	0.09
5	0.24	1.00	0.52
6	0.07	0.44	0.15

- That 0.005% of persons with negative test results are suffering from the disease
- That 99.995% of 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) do 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. Now, let us consider a little more sophisticated example, shown in Table IV. Attributes disease, age, and sex are condition attributes, whereas test is the decision attribute.

The strength, certainty, and coverage factors for the decision table are shown in Table V. The flow graph for the decision table is presented in Fig. 1.

The following is a decision algorithm associated with Table IV:

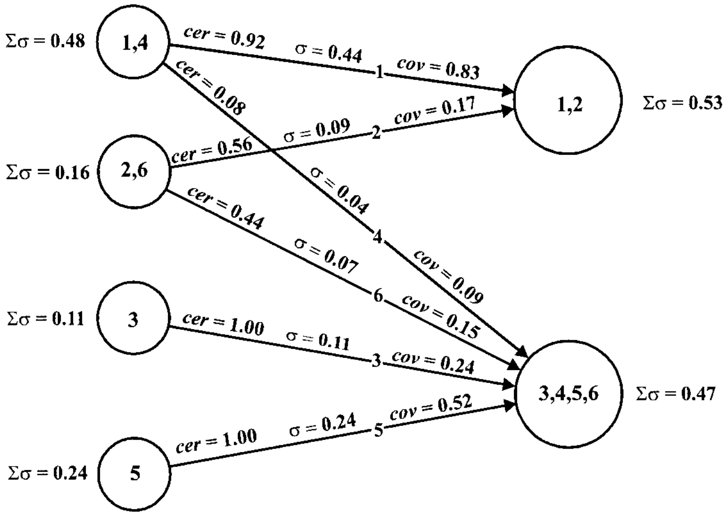

Figure 1. Flow graph.

Table VI. 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

(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 foregoing algorithm are given in Table VI. The flow graph for the decision algorithm is presented in Fig. 2.

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

- Ninety-two percent of ill and old patients have positive test results
- Fifty-six percent of ill and middle-aged patients have positive test results
- All healthy patients have negative test results
- Eight percent of ill and old patients have negative test results
- Forty-four percent of ill and middle-aged patients have negative test results

In other words:

- Ill and old patients most probably have positive test results (probability = 0.92)
- Ill and middle-aged patients most probably have positive test results (probability = 0.56)
- Healthy patients have certainly negative test results (probability = 1.00)

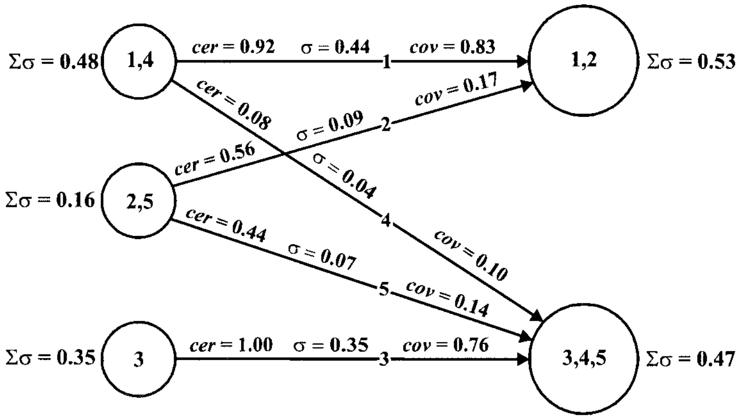

Figure 2. Flow graph for the decision algorithm.

Now let us examine the following 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)

Using 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)

■

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

8. CONCLUSION

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

In the rough set approach, Bayes' theorem reveals data patterns, used next to draw conclusions from data, in the form of decision rules, which refer to objective probabilities computed from data. Furthermore, the proposed approach enables us to apply Bayes' theorem to give reasons for decisions by using reversed decision rules. Moreover, the association of flow graphs with decision tables gives new dimension to decision process analysis.

Acknowledgments

The authors thank anonymous referees for critical comments.

References

1. Pawlak Z. Logic, probability, and rough sets. In: Karhumaki J, Maurer H, Paun G, Rozenberg G, editors. Jewels are forever, contributions on theoretical computer science in honor of Arto Salomaa. Berlin: Springer-Verlag; 1999. pp 364–373.
2. Łukasiewicz J. Die logishen Grundlagen der Wahrscheinilchkeitsrechnung. Kraków, 1913. In: Borkowski L, editor. Jan Łukasiewicz—selected works. Amsterdam: North Holland Publishing Co.; Warsaw: Polish Scientific Publishers; 1970. pp 16–63.
3. Bayes T. An essay toward solving a problem in the doctrine of chances. Philos Trans R Soc 1763;53:370–418; Reprint Biometrika 1958;45:296–315.
4. Bernardo JM, Smith AFM. Baysian theory. New York: Wiley; 1994.

5. Berthold M, Hand DJ. Intelligent data analysis. An introduction. Berlin: Springer-Verlag; 1999.
6. Box GEP, Tiao GC. Bayesian inference in statistical analysis. New York: Wiley; 1992.
7. Adams EW. The logic of conditionals, an application of probability to deductive logic. Dordrecht: D. Reidel Publishing Co.; 1975.
8. Pawlak Z. Rough sets, decision algorithms and Bayes' theorem. Eur J Oper Res 2002; 136:181–189.
9. Pawlak Z, Skowron A. Rough membership functions. In: Yager R, Fedrizzi M, Kacprzyk J, editors. Advances in the Dempster-Shafer theory of evidence. New York: Wiley; 1994. pp 251–271.
10. Tsumoto S, Tanaka H. Discovery of functional components of proteins based on PRIMEROSE and domain knowledge hierarchy. Soft Computing, San Jose, CA, 1995. pp 780–785.
11. Polkowski L, Skowron A, editors. Rough sets in knowledge discovery; Heidelberg: Physica-Verlag; 1998. 1:1–601; 2:1–576.
12. Ford LR, Fulkerson DR. Flows in networks. Princeton, NJ: Princeton University Press; 1962.