Using Rough Sets for Drawing Conclusions from Data

Zdzisław Pawlak

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

Summary. Rough set based data analysis starts from a data table, called an *information system*. The information system contains data about objects of interest characterized in terms of some attributes. Often we distinguish in the information system condition and decision attributes. Such information system is called a *decision table*. The decision table describes decisions in terms of conditions that must be satisfied in order to carry out the decision specified in the decision table. With every decision table a set of decision rules, called a *decision algorithm* can be associated. It is shown that every decision algorithm reveals some well known probabilistic properties, in particular it satisfies the total probability theorem and the Bayes' theorem. These properties give a new method of drawing conclusions from data, without referring to *prior* and *posterior* probabilities, inherently associated with Bayesian reasoning.

Kaywords: Rough sets, decision tables, decision algorithm, Bayes' Theorem.

1 Introduction

The paper concerns a relationship between rough sets and the Bayes' theorem. It reveals new look on the Bayes' theorem from the rough set perspective and is continuation of ideas presented in [5, 6].

In the paper basic notions of rough set theory will be given and the notion of the decision algorithm will be defined and some its basic properties will be shown. It is revealed that every decision table (decision algorithm) displays well known probabilistic features, in particular it satisfies the total probability theorem and the Bayes' theorem. These properties give a new method of drawing conclusions from data, without referring to prior and posterior probabilities, inherently associated with Bayesian reasoning.

The revealed relationship can be used to invert decision rules, i.e., giving reasons (explanations) for decisions, which can be very useful in decision analysis.

Summing up statistical inference based on Bayes' theorem is used to verify prior knowledge when the data become available, whereas rough set inference based on Bayes' theorem uses relationships in the data revealed by Bayes' theorem.

In other words, rough set view on Bayes' theorem reveals relationship between conditions and decisions in decision rules, uses strength of decision rules is a basis for computation and reveals relationships in any decision table without referring either to prior or posterior probabilities.

Basic of rough set theory can be found in [2, 4]. More advanced topics are discussed in [7, 8].

2 Approximation of sets

Starting point of rough set based data analysis is a data set, called an information system.

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.

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 Decision rules

In what follows we will define a formal language to describe decision tables in logical terms.

Let $S = (U, A)$ be an information system. With every $B \subseteq A$ we associate a formal language, i.e., a set of formulas $For(B)$. Formulas of $For(B)$ are built up from attribute-value pairs (a, v) where $a \in B$ and $v \in V_a$ by means of logical connectives $\wedge$ (*and*), $\vee$ (*or*), $\sim$ (*not*) in the standard way.

For any $\Phi \in For(B)$ by $||\Phi||_S$ we denote the set of all objects $x \in U$ satisfying Φ in S and refer to as the *meaning* of Φ in S.

The meaning $||\Phi||_S$ of Φ in S is defined inductively as follows:
$||(a, v)||_S = \{x \in U : a(v) = x\}$ for all $a \in B$ and $v \in V_a$, $||\Phi \vee \Psi||_S = ||\Phi||_S \cup ||\Psi||_S, ||\Phi \wedge \Psi||_S = ||\Phi||_S \cap ||\Psi||_S, || \sim \Phi||_S = U - ||\Phi||_S$.

If $S = (U, C, D)$ is a decision table then with every row of the decision table we associate a decision rule, which is defined next.

A *decision rule* in S is an expression $\Phi \rightarrow_S \Psi$, or simply $\Phi \rightarrow \Psi$ if S is understood, read *if* Φ *then* Ψ, where $\Phi \in For(C)$, $\Psi \in For(D)$ and C, D are condition and decision attributes, respectively; Φ and Ψ are referred to as *conditions* and *decisions* of the rule, respectively.

The number $supp_S(\Phi, \Psi) = card(||\Phi \wedge \Psi||_S)$ will be called the *support* of the rule $\Phi \rightarrow \Psi$ in S. We consider a probability distribution $p_U(x) = 1/card(U)$ for $x \in U$ where U is the (non-empty) universe of objects of S; we have $p_U(X) = card(X)/card(U)$ for $X \subseteq U$. For any formula Φ we associate its probability in S defined by

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

With every decision rule $\Phi \rightarrow \Psi$ we associate a conditional probability

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

called the *certainty factor* of the decision rule, denoted $cer_S(\Phi, \Psi)$. This idea was used first by Łukasiewicz [3] (see also [1]) to estimate the probability of implications. We have

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

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

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

Obviously, $\pi_S(\Psi|\Phi) = 1$ if and only if $\Phi \to \Psi$ is true in S.

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

There is an interesting relationship between decision rules and approximations: certain decision rules correspond to the lower approximation, whereas the uncertain decision rules correspond to the boundary region. More about this relationships can be found in [8].

Besides, we will also use a *coverage factor* of the decision rule, denoted $cov_S(\Phi, \Psi)$ (used e.g. by Tsumoto and Tanaka [10] for estimation of the quality of decision rules) defined by

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

Obviously we have

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

There are three possibilities to interpret the certainty and the coverage factors: statistical (frequency), logical (degree of truth) and mereological (degree of inclusion).

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

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

The number

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

will be called the *strength* of the decision rule $\Phi \to \Psi$ in S, and will play an important role in our approach, which will be discussed in section 6.

We will need also the notion of an equivalence of formulas.

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

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

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

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

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

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

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

4 Decision algorithms

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

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

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

The set of decision rules $Dec(S)$ that satisfies 1), 2) 3) and 4), i.e., is independent, covers U, preserves the consistency of S and all decision rules $\Phi \to \Psi \in Dec(S)$ are admissible in S – will be called a *decision algorithm* in S.

Hence, if $Dec(S)$ is a decision algorithm in S then the conditions of rules from $Dec(S)$ define in S a partition of U. Moreover, the *positive region of D with respect to C*, i.e., the set

$$\bigcup_{X \in U/D} C_*(X)$$

is partitioned by the conditions of some of these rules, which are certain in S.

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

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

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

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

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

The number

$$\eta(Dec(S)) = \sum_{\Phi\to\Psi\in Dec(S)} max\{\sigma_S(\Phi,\Psi)\}_{\Psi\in D(\Phi)}$$

where $D(\Phi) = \{\Psi : \Phi \to \Psi \in Dec(S)\}$ will be referred to as the *efficiency* of the decision algorithm $Dec(S)$ in S, and the sum is stretching over all decision rules in the algorithm.

The efficiency of a decision algorithm is the probability (ratio) of all objects of the universe, that are classified to decision classes, by means of decision rules $\Phi \to \Psi$ with maximal strength $\sigma_S(\Phi,\Psi)$ among rules $\Phi \to \Psi \in Dec(S)$ with satisfied Φ on these objects. In other words, the efficiency says how well the decision algorithm classifies objects when the decision rules with maximal strength are used only.

As mentioned at the beginning of this section decision algorithm is a counterpart of a decision table. The properties 1) - 4) have been chosen in such a way that the decision algorithm preserves basic properties of the data in the decision table, in particular approximations and boundary regions of decisions.

Crucial issue in the rough set based data analysis is the generation of "optimal" decision algorithms from the data. This is a complex task, particularly when large data bases are concerned. Many methods and algorithms have been proposed to deal with this problem but we will not dwell upon this issue here, for we intend restrict this paper to rudiments of rough set theory only. The interested reader is advised to consult the references [7, 8] and the web.

5 Decision algorithms and approximations

Decision algorithms can be used as a formal language for describing approximations.

Let $Dec(S)$ be a decision algorithm in S and let $\Phi \to \Psi \in Dec(S)$. By $C(\Psi)$ we denote the set of all conditions of Ψ in $Dec(S)$ and by $D(\Phi)$ - the set of all decisions of Φ in $Dec(S)$.

Then we have the following relationships:

a) $C_*(||\Psi||_S) = || \bigvee_{\Phi'\in C(\Psi),\ \pi(\Psi|\Phi')=1} \Phi'||_S,$

b) $C^*(||\Psi||_S) = || \bigvee_{\Phi'\in C(\Psi),\ 0<\pi(\Psi|\Phi')\leq 1} \Phi'||_S,$

c) $BN_C(||\Psi||_S) = || \bigvee_{\Phi'\in C(\Psi),\ 0<\pi(\Psi|\Phi')<1} \Phi'||_S.$

From the above properties we can get the following definitions:

i) If $||\Phi||_S = C_*(||\Psi||_S)$, then formula Φ will be called the *C-lower approximation* of the formula Ψ and will be denoted by $C_*(\Psi)$;
ii) If $||\Phi||_S = C^*(||\Psi||_S)$, then the formula Φ will be called the *C-upper approximation* of the formula Φ and will be denoted by $C^*(\Psi)$;
iii) If $||\Phi||_S = BN_C(||\Psi||_S)$, then Φ will be called the *C-boundary* of the formula Ψ and will be denoted by $BN_C(\Psi)$.

The above properties say that any decision $\Psi \in Dec(S)$ can be uniquely described by the following certain and uncertain decision rules respectively:

$$C_*(\Psi) \to \Psi,$$

$$BN_C(\Psi) \to \Psi.$$

This property is an extension of some ideas given by Ziarko [11].

6 Some properties of decision algorithms

Decision algorithms have interesting probabilistic properties which are discussed in this section.

Let $Dec(S)$ be a decision algorithm and let $\Phi \to \Psi \in Dec(S)$. Then the following properties are valid:

$$\sum_{\Phi' \in C(\Psi)} cer_S(\Phi', \Psi) = 1 \tag{1}$$

$$\sum_{\Psi' \in D(\Phi)} cov_S(\Phi, \Psi') = 1 \tag{2}$$

$$\pi_S(\Psi) = \sum_{\Phi' \in C(\Psi)} cer_S(\Phi', \Psi) \cdot \pi_S(\Phi') = \sum_{\Phi' \in C(\Psi)} \sigma_S(\Phi', \Psi) \tag{3}$$

$$\pi_S(\Phi) = \sum_{\Psi' \in D(\Phi)} cov_S(\Phi, \Psi') \cdot \pi_S(\Psi') = \sum_{\Psi' \in D(\Phi)} \sigma_S(\Phi, \Psi') \tag{4}$$

$$\begin{aligned} cer_S(\Phi, \Psi) &= \frac{cov_S(\Phi, \Psi) \cdot \pi_S(\Psi)}{\sum_{\Psi' \in D(\Phi)} cov_S(\Phi, \Psi') \cdot \pi_S(\Psi')} = \\ &= \frac{\sigma_S(\Phi, \Psi)}{\sum\limits_{\Psi' \in D(\Phi)} \sigma_S(\Phi, \Psi')} = \frac{\sigma_S(\Psi, \Phi)}{\pi_S(\Phi)} \end{aligned} \tag{5}$$

$$\begin{aligned} cov_S(\Phi, \Psi) &= \frac{cer_S(\Phi, \Psi) \cdot \pi_S(\Phi)}{\sum_{\Phi' \in C(\Psi)} cer_S(\Phi', \Psi) \cdot \pi_S(\Phi')} = \\ &= \frac{\sigma_S(\Phi, \Psi)}{\sum\limits_{\Phi' \in C(\Psi)} \sigma_S(\Phi', \Psi)} = \frac{\sigma_S(\Phi, \Psi)}{\pi_S(\Psi)} \end{aligned} \tag{6}$$

That is, any decision algorithm, and consequently 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 the *Bayes' theorem*. Note that we are not using to prior and posterior probabilities – fundamental in Bayesian data analysis philosophy.

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 in the decision algorithm only. The strength of decision rules can be computed from the data or can be a subjective assessment.

In other words, if we know the ratio of Φ_S in Ψ, thanks to the Bayes' theorem, we can compute the ratio of Ψ_S in Φ.

References

1. Adams, E. W. (1975) The Logic of Conditionals, an Application of Probability to Deductive Logic. D. Reidel Publishing Company, Dordrecht, Boston
2. Düntsch, I.Gediga, G. (2000) Rough Set Data Analysis – a Road to Non-invasive Knowledge Discovery. Metoδos Publisher, Bangor, Bissendorf
3. Łukasiewicz, J. Die logishen Grundlagen der Wahrscheinilchkeitsrechnung. Krakow, 1913. In: L. Borkowski (ed.), Jan Łukasiewicz – Selected Works, North Holland Publishing Company, Amsterdam, London, Polish Scientific Publishers, Warsaw, 1970
4. Pawlak, Z. (1991) Rough Sets – Theoretical Aspects of Reasoning about Data; Kluwer Academic Publishers: Boston, Dordrecht
5. Pawlak, Z.(1999) Decision Rules, Bayes' Rule and Rough Sets. In: N. Zhong, A. Skowron, S. Ohsuga (eds.), Proceedings of 7th International Workshop: New Directions in Rough Sets, Data Mining, and Granular –Soft Computing (RSFDGSC'99), Yamaguchi, Japan, November 1999, Lecture Notes in Artificial Intelligence 1711 Springer–Verlag, Berlin, 1–9
6. Pawlak, Z. (2000) Rough Sets and Decision Algorithms. Springer (to appear)
7. Polkowski, L., Skowron, A. (1998) Proceedings of the First International Conference Rough Sets and Current Trends in Computing (RSCTC'98), Warsaw, Poland, June, Lecture Notes in Artificial Intelligence 1424, Springer–Verlag, Berlin
8. Polkowski, L., Skowron, A. (1998) Rough Sets in Knowledge Discovery Vol. 1–2, Physica-Verlag, Heidelberg
9. A. Skowron, 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, pp. 1-17, 2000.
10. S. Tsumoto and H. Tanaka, 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, pp. 280-285, SCS, 1995.
11. Ziarko, W. (1998) Approximation region-based decision tables. In: L. Polkowski, A. Skowron (eds.), Rough Sets in Knowledge Discovery Vol.1-2, Physica-Verlag, Heidelberg, 178–185