Theorize with Data using Rough Sets

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

Abstract

In this paper we will study the granular structure of data in the language of rough set theory. It is shown that the granularity of data can be represented in a form of a flow graph, and the relationship between granules obeys Bayes' theorem. This leads to a new method of data analysis.

Motto:
„It is a capital mistake to theorise before one has data"
Sherlock Holmes
In: A Scandal in Bohemia

This paper is dedicated to the renowned Mr. Sherlock Holmes for his mastery in theorizing about data.

1. Introduction

In the rough set approach to data analysis, patterns in data are characterized by means of approximations, or equivalently by decision rules induced by the data, presented in a form of a decision table. With every decision rule in a decision table three coefficients are associated: the *strength*, the *certainty* and the *coverage factors* of the rule. These coefficients satisfy Bayes' theorem and the total probability theorem. This enables us to use Bayes' theorem to discover patterns in data in a different way from that offered by standard Bayesian inference technique without referring to prior and posterior probabilities, intrinsically embedded in Bayesian inference methodology.

In the presented approach granularity imposed by the indiscernibility relation on data plays an important role. With every decision table a flow graph is associated, which defines a connection between condition and decision granules in the decision table. The certainty and coverage factors determine granular structure of data, which shows the relationship between condition and decision granules determined by the decision table.

The granular structure of data can be represented in a Euclidean „decision space", in which dimensions are determined by decision granules, points in the space are condition granules and coordinates of the points are strengths of the corresponding rules. Distance in the decision space between condition granules allows to determine how „distant" are decision granules in terms of data.

A simple tutorial example will be used to illustrate the basic ideas of the presented approach.

2. Basic concepts

In this section we define basic concepts of rough set theory.

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, an *information system* is a pair $S = (U,A)$, where U and A, are non-empty finite sets called the *universe*, and the set of *attributes*, respectively such that $a:U \rightarrow V_a$, where V_a, is the set of all *values* of a 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 $B(x)$ and called B-granule induced by 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 the 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 and $C \cup D = A$.

$C(x)$ and $D(x)$ will be referred to as the condition granule and the decision granule induced by x, respectively.

3. Decision rules

Every row of a decision table determines a decision rule.

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 $C \xrightarrow{x} D$.

The number $supp_x(C,D) = |A(x)| = |C(x) \cap D(x)|$ will be called a *support* of the decision rule $C \xrightarrow{x} D$ and the number

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

will be referred to as the *strength* of the decision rule $C \xrightarrow{x} D$, where $|X|$ denotes the cardinality of X.

With every decision rule $C \xrightarrow{x} D$ we associate a *certainty factor* of the decision rule, denoted $cer_x(C, D)$ and defined as follows:

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

where $C(x) \neq \varnothing$ and $\pi(C(x)) = \frac{|C(x)|}{|U|}$.

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

If $cer_x(C, D) = 1$, then $C \xrightarrow{x} D$ will be called a *certain decision* rule; if $0 < cer_x(C, D) < 1$ the decision rule will be referred to as an *uncertain decision rule*.

Besides, we will also use a *coverage factor* (see [5]) of the decision rule, denoted $cov_x(C, D)$ defined as

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

where $D(x) \neq \varnothing$ and $\pi(D(x)) = \frac{|D(x)|}{|U|}$.

Similarly

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

If $C \xrightarrow{x} D$ is a decision rule then $D \xrightarrow{x} C$ will be called a *inverse decision rule*. The inverse decision rules can be used to give *explanations* (*reasons*) for a decision.

4. Properties of decision rules

Decision rules have important probabilistic properties which are discussed next.

Let $C \xrightarrow{x} D$ be a decision rule. Then the following properties are valid:

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

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

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

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

$$cer_x(C,D) = \frac{cov_x(C,D) \cdot \pi(D(x))}{\pi(C(x))} = \frac{\sigma_x(C,D)}{\pi(C(x))} \tag{5}$$

$$cov_x(C,D) = \frac{cer_x(C,D) \cdot \pi(C(x))}{\pi(D(y))} = \frac{\sigma_x(C,D)}{\pi(D(x))} \tag{6}$$

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 formula (5) and (6) it is enough to know the strength (support) of all decision rules only.

Formulas (5) and (6) can be rewritten as

$$cer_x(C,D) = cov_x(C,D) \cdot \gamma_x(C,D) \tag{7}$$

$$cov_x(C,D) = cer_x(C,D) \cdot \gamma_x^{-1}(C,D) \tag{8}$$

where $\gamma_x(C,D) = \frac{|D(x)|}{|C(x)|} = \frac{cer_x(C,D)}{cov_x(C,D)}$

called a *granularity factor* of the decision rule induced by x, reveals the granular structure of the decision rule.

Besides, the granularity factor exhibits the granular structure of Bayes' theorem and thus enables us to connect Bayes' theorem with granular structure of data.

5. Granularity of data 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 $C \xrightarrow{x} D$ 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).

The application of flow graphs to represent relationship between data granules gives a clear insight into the granular structure of data analysis process. Classification of objects in this representation boils down to finding the maximal output flow in the flow graph, whereas explanation of decisions is connected with the maximal input flow associated with the given decision.

6. Decision space and granularity of data

With every decision table having one *n*-valued decision attribute we can associate *n*-dimensional Euclidean space, where decision granules determine *n* axis of the space and condition granules determine points of the space. Strengths of decision rules are to be understood as coordinates of corresponding granules.

Distance $\delta(x, y)$ between granules *x* and *y* in an *n*-dimensional decision space is defined as

$$\delta(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

where $x = (x_1,...,x_n)$ and $y = (y_1,...,y_n)$ are vectors of strengths of corresponding decision rules.

It follows from the above example that granules B, C and D are „close" and form a cluster which is „distant" from granule A.

7. An example

An example of a simple decision table is shown below.

Table 1. Decision table

decision rule	*age*	*sex*	*profession*	*disease*
1	*old*	*male*	*yes*	*no*
2	*med.*	*female*	*no*	*yes*
3	*med.*	*male*	*yes*	*no*
4	*old*	*male*	*yes*	*yes*
5	*young*	*male*	*no*	*no*
6	*med.*	*female*	*no*	*no*

In the table *age*, *sex* and *profession* are condition attributes, whereas *disease* is the decision attribute.

The table contains data concerning relationship between age, sex, profession and certain vocational disease.

In Table 2 below a modified version of Table 1 is shown.

Table 2. Support and strength

d. rule	*age*	*sex*	*prof.*	*disease*	*supp.*	*stren.*
1	*old*	*male*	*yes*	*no*	200	0.18
2	*med.*	*female*	*no*	*yes*	70	0.06
3	*med.*	*male*	*yes*	*no*	250	0.23
4	*old*	*male*	*yes*	*yes*	450	0.41
5	*young*	*male*	*no*	*no*	30	0.03
6	*med.*	*female*	*no*	*no*	100	0.09

Certainty and coverage factors for the decision table presented in Table 2 are shown in Table 3.

Table 3. Certainty and coverage factors

decision rule	*strength*	*certainty*	*coverage*
1	0.18	0.31	0.34
2	0.06	0.40	0.13
3	0.23	1.00	0.43
4	0.41	0.69	0.87
5	0.03	1.00	0.06
6	0.09	0.60	0.17

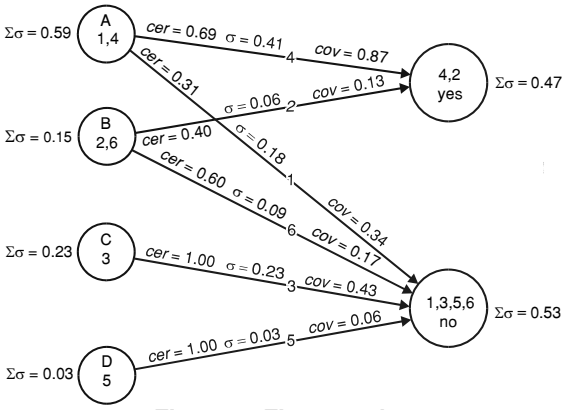

Figure 1. Flow graph

Flow graph associated with decision table presented in Table 2 is shown in Fig. 1.

Decision space for Table 1 is shown in Figure 2

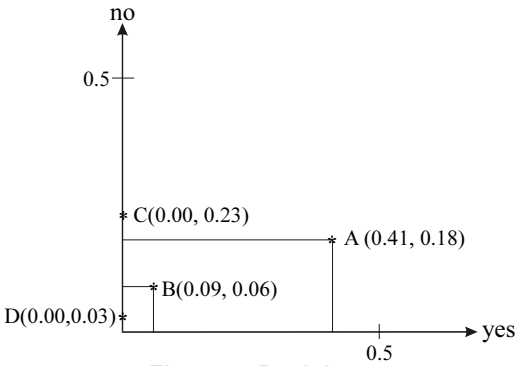

Figure 2. Decision space

Distances between granules A, B, C and D are shown in Table 4.

	A	B	C	D
A				
B	0.4511			
C	0.4130	0.1523		
D	0.4365	0.0849	0.2000	

Table 4. Distance matrix

8. Conclusions

The indiscernibility relation, the basic concept of rough set theory imposes granularity as a fundamental paradigm in data analysis.

It is shown in this paper a relationship between rough set theory, and Bayes' theorem in terms of granularity of data. Besides, the relation between condition and decision granules is represented as a flow graph. Moreover, a decision space is defined to analyze similarity of data granules.

9. References

[1] T.Y. Lin, "Granular Computing on Binary Relations, Data Mining and Neighborhood Systems", L. Polkowski, A. Skowron (eds.), Rough Sets in Knowledge Discovery 1. Methodology and Applications, Physica-Verlag, Springer Verlag Company, Heidelberg, 1998, pp. 107-121.

[2] Z. Pawlak, „Granularity, Multi-valued Logic, Bayes' Theorem and Rough Sets", Data Mining Rough Sets and Granular Computing, Tsan Young Lin, Yiyu Y. Yao, Lotti A. Zadeh (eds.), Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer Verlag Company, 2002, pp. 487-498.

[3] Z. Pawlak, "Granularity of Knowledge, Indiscernibility and Rough Sets", IEEE International Conference on Granulationary Computing, May 5-9, Anchorage, Alaska, 1998, pp. 100-103.

[4] Z. Pawlak, "New Look on Bayes' Theorem - the Rough Set Outlook", Proceedings 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, pp. 1-8.

[5] L. Polkowski, A. Skowron, "Calculi of Granules Based on Rough Set Theory: Approximate Distributed Synthesis and Granular Semantics for Computing with Words", Lecture Notes in Artificial Intelligence 1711, N. Zhong, A. Skowron, S. Ohsuga (eds.), New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, 7th International Workshop, RSFDGrC'99, Yamaguchi, Japan, November 1999, pp. 20-28.

[6] L. Polkowski, A. Skowron, "Towards Adaptative Calculus of Granules", Proceedings of the FUZZY-IEEE'98 International Conference, Anchorage, Alaska, USA, May 5-9, 1998.

[7] A. Skowron, J. Stepaniuk, "Information Granules in Distributed Environment", Lecture Notes in Artificial Intelligence 1711, N. Zhong, A. Skowron, S. Ohsuga (eds.), New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, 7th International Workshop, RSFDGrC'99, Yamaguchi, Japan, November 1999, pp. 357-364.

[8] A. Skowron, J. Stepaniuk, *Information Granulation – a Rough Set Approach,* Manuscript, 1997.

[9] A. Skowron, J. Stepaniuk, "Information Granules and Approximation Spaces", Proceedings of the 7th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems, IPMU'98, Paris, France, July 6-10, 1998, pp. 1354-1361.

[10] A. Skowron, J. Stepaniuk, J.F. Peters, "Approximation of Information Granules Sets", Rough Sets and Current Trends in Computing, LNAI 2005, W. Ziarko, Y.Yao (eds.), Second International Conference, RSCTC 2000, Banff, Canada, October 2000, pp. 65-73.

[11] S. Tsumoto, 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, SCS, 1995, pp. 280-285.

[12] T.Y. Lin, Y.Y. Yao, L.A. Zadeh (eds.), *Data Mining Rough Sets and Granular Computing,* Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer Verlag Company, 2002.

[13] Y.Yamauchi, M.Mukaidono, "Probabilistic Inference and Bayesian Theorem on Rough Sets", Rough Sets and Current Trends in Computing LNAI 2005, W. Ziarko, Y.Yao (eds.), Second International Conference, RSCTC 2000, Banff, Canada, October 2000, pp. 73-81

[14] W. Ziarko, Y.Yao (eds.), *Rough Sets and Current Trends in Computing*, LNAI 2005, Second International Conference, RSCTC 2000, Banff, Canada, October 2000.

[15] L. Zadeh, *Fuzzy Graphs, Rough Sets and Information Granularity*, Proceedings Third Int. Workshop on Rough Sets and Soft Computing, November 10-12, San Jose, 1994.

[16] L. Zadeh, *The Key Rules of Information Granulation and Fuzzy Logic in Human Reasoning, Concept Formulation and Computing with Words*, Proceedings FUZZ-96: Fifth IEEE International Conference on Fuzzy Systems, September 8-11, New Orleans, 1996.

[17] L. Zadeh, *Information Granulation, Fuzzy Logic and Rough Sets*, Proceedings of the Fourth Int. Workshop on Rough Sets, and Machine Discovery, November 6-8, Tokyo, 1996.

[18] N. Zhong, A. Skowron, S. Ohsuga (eds.), *New Directions in Rough Sets, Data Mining, and Granular-Soft Computing*, Lecture Notes in Artificial Intelligence 1711, 7th International Workshop, RSFDGrC'99 Yamaguchi, Japan, November 1999.