# ROUGH SETS THEORY AND DATA REDUCTION IN INFORMATION SYSTEMS AND DATA MINING

**Mofreh Hogo, Miroslav Šnorek**

CTU in Prague, Departement Of Computer Sciences And Engineering Karlovo Náměstí 13, 121 35 Prague 2,CR, E-mail: mofreh_hogo@hotmail.com

**Abstract**

The major aims of this work is to introduce an excellent method for data reduction in information systems and also in data mining, based on rough sets theory, this method capable of discovering the relevant features or relevant attributes which are useful for data description and/or prediction, and to filter out the irrelevant ones. As well as data reduction step is considered as very important step in data mining or knowledge discovering process. Where the aim of the data reduction phase is to decrease the computation (learning time) effort for inducing efficient classifier (major ask of the data mining tasks) as well as producing simple structure of this classifier. The developed system is tested with two data sets and it was proved its ability for decreasing the complexity required, as well as simplifying the structure for the built classifiers.

**Keyword:** Data Reduction, Data Mining, Information System, Rough Sets Theory, Rule Generation

## 1. Introduction

We are much better at collecting data than we are at using it in a sensible way, and the amounts we collect outstrip our ability to use them with existing methods. The term *data mining* was first used in the late 80s to redress the balance between collecting and understanding data. Data mining is defined as the process of discovering interesting knowledge: patterns, associations, trends, shifts and anomalies from large amounts of data stored in databases, data warehouses, or other information repositories. Lewinson (1994) defines Data Mining as "analyzing historical data to find patterns that shed light on the present". Due to the wide availability of huge amounts of data in electronic form and the need for turning such data into useful information and knowledge, data mining has attracted a great deal of attention in the information industry in recent years ( Frawley et al. 1992;Fayyad et al. 1996; Shapiro et al. 1996). The field has far-reaching applications including market analysis, advance diagnosis, business management and decision support. Data mining had been popularly treated as a synonym of *knowledge discovery* in databases, although some researchers view data mining as an essential step of knowledge discovery. Fig.1 shows the major steps of KDD. In general, a knowledge discovery process consists of an interactive sequence of the following steps[1]:

1. *Data cleaning*, which handles noisy, erroneous, missing, or irrelevant data,
2. *Data integration*, where multiple, heterogeneous data sources may be integrated into one,
3. *Data selection*, where data relevant to the analysis task are retrieved from the database,
4. *Data transformation*, where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations,

5. *Data mining*, which is an essential process where intelligent methods are applied in order to extract data patterns,
6. *Pattern evaluation*, which is to identify the truly interesting patterns representing kno-dge based on some "interestingness" measures, and
7. *Knowledge presentation*, where visualization and knowledge presentation techniques are used to present the mined knowledge to the user.

In this work we have focused only on step 3 which is, "data selection". Our paper is structured as follows. In Section 2 below we present some necessary basic facts about the rough set theory. Then in Section 3 we introduce a short example to expalin the ability of rough sets theory for attributes selection. Section 4 introduces the applied algorithm. Section 5 presents the used data sets and the experimental results. Section 6 the conclusion.

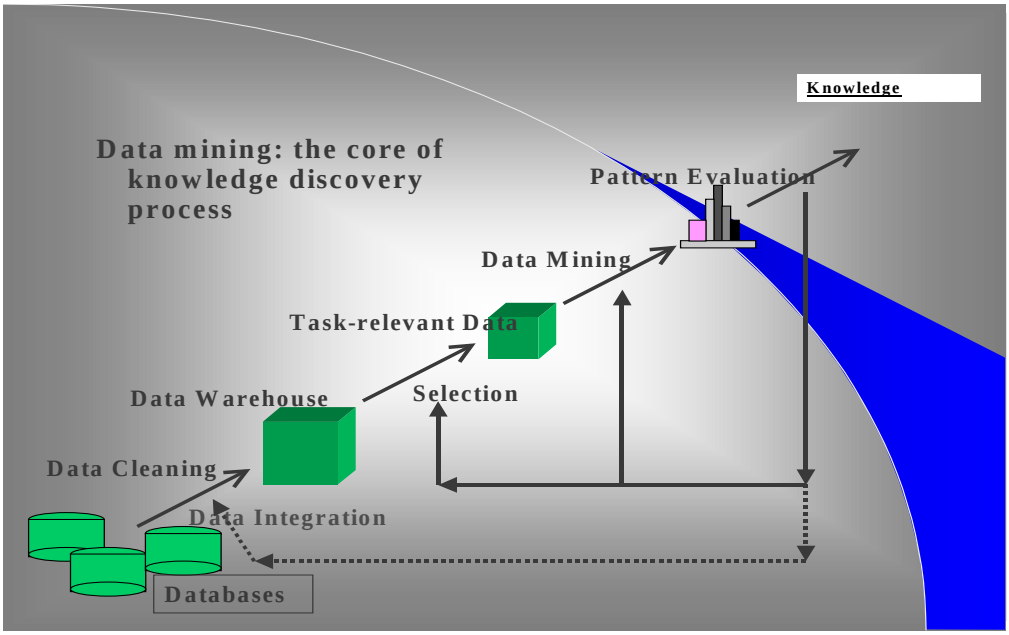


Fig.1: An Overview Of The Steps Comprising The KDD Process

## 2. Basic rough sets theory

### *2.1 Information Systems*

Information systems [2][3] (sometimes called data tables, attribute-value systems, condition-action tables, knowledge representation systems etc.) are used for representing knowledge. Rough sets have been introduced as a tool to deal with inexact, uncertain or vague knowledge in artificial intelligence applications. In this section we recall some basic notions related to information systems and rough sets.

An *information system* is a pair A = (*U*, *A*), where *U* is a non-empty, finite set called the *universe* and *A* is a non-empty, finite set of *attributes*, i.e. $a$: $U \to V_a$ for $a \in A$, where $V_a$ is called the *value set* of $a$. Elements of *U* are called *objects* and interpreted as, e.g. cases, states, processes, patients, observations. Attributes are interpreted as features, variables, characteristic conditions etc. Every information system A = (*U*, *A*) and non-empty set $B \subseteq A$ determine a *B-information function* $Inf_B : U \to P(B \times \bigcup_{a \in B} V_a)$ defined by $Inf_B(x)=\{(a,a(x)): a \in B\}$.

The set$\{Inf_A(x):x\in U\}$is called the *A-information set* and it is denoted by INF(A). With every subset of attributes $B\subseteq A$, an equivalence relation, denoted by $IND_A(B)$ (or $IND(B)$) called the *B-indiscernibility relation*, is associated and defined by $IND(B)=\{(s,s')\in U^2$: for every $a\in B$, $a(s)=a(s')\}$. Objects $s,s'$ satisfying relation $IND(B)$ are *indiscernible* by attributes from $B$. Hence $xIND(A)y$ iff $Inf_A(x)=Inf_A(y)$. We consider a special case of information systems called *decision tables.* A decision table is any information system of the form A= ($U$, $A\cup\{d\}$), where $d\in A$ is a distinguished attribute called the *decision*. The elements of $A$ are called *conditions*. One can interpret a decision attribute as a kind of classification of the universe of objects given by an expert, decision-maker, operator, physician, etc. Decision tables are called *training sets of examples* in machine learning. The cardinality of the image $d(U)=\{k: d(s)=k$ for some $s\in U\}$ is called the *rank* of $d$ and is denoted by $r(d)$. We assume that the set $V_d$ of values of the decision $d$ is equal to $\{1,...,r(d)\}$. Let us observe that the decision $d$ determines the partition $CLASS_A(d)=\{X_1,...,X_{r(d)}\}$ of the universe $U$, where $X_k=\{x\in U: d(x)=k\}$ for $1\le k\le r(d)$. $CLASS_A(d)$ will be called *the classification of objects in* A *determined by the decision d*. The set $X_i$ is called the *i-th decision class of* A.

### 2.2 Reducts

Any minimal subset $B\subseteq A$ such that $IND(A)=IND(B)$ is called *a reduct in the information system* A. The set of all reducts in A is denoted by $RED$(A). Let A be an information system with $n$ objects. By $M$(A), we denote an $n\times n$ matrix $(c_{ij})$ called the *discernibility matrix* of A such that, $c_{ij}=\{a\in A: a(x_i)\neq a(x_j)\}$ for $i,j=1,...,n$. A *discernibility function* $f_A$ for an information system A is a boolean function of $m$ boolean variables $\bar{a}_1,..,\bar{a}_m$ corresponding to the attributes $a_1,...,a_m$ respectively, and defined by $f_A(\bar{a}_1,..,\bar{a}_m)=\wedge\{\vee\bar{c}_{ij}: 1\le j<i\le n, c_{ij}\neq\emptyset\}$, where $\bar{c}_{ij}=\{\bar{a}: a\in c_{ij}\}$. It can be shown that the set of all prime implicants of $f_A$ determines the set $RED$(A) of all reducts of A (i.e. $a_{i_1}\wedge...\wedge a_{i_k}$ is a prime implicant of $f_A$. iff $\{a_{i_1},\ldots,a_{i_k}\}\in RED$(A). One can show that the problem of minimal (with respect to the cardinality) reduct finding is NP-hard. In general the number of reducts of a given information system can be exponential with respect to the number of attributes. Nevertheless, existing procedures for reduct computation are efficient in many practical applications and for more complex cases one can apply some efficient heuristics.

### 2.3 Set Approximations

If A = ($U$, $A$) is an information system[4], $B\in A$ is a set of attributes and $X\subseteq U$ is a set of objects then the sets $\{s\in U: [s]_B\subseteq X\}$ and $\{s\in U: [s]_B\cap X\neq\emptyset\}$ are called *B-lower and B-upper approximation* of $X$ in A, and they are denoted by $\underline{B}X$ and $\overline{B}X$, respectively. The set $BN_B(X)=\underline{B}X-\overline{B}X$, will be called the *B-boundary* of $X$. These concepts will be shown in Fig. 1. When $B=A$ we write also $BN_A(X)$ instead of $BN(X)$. Sets which are unions of some classes of the indiscernibility relation $IND(B)$ are called definable by $B$. The set $X$ is *B-definable* iff $\underline{B}X=\overline{B}X$. Some subsets (categories) of objects in an information system cannot be expressed exactly by employing available attributes but they can be roughly defined. The set $\underline{B}X$ is the set of all elements of $U$ which can be with certainty classified as elements of $X$, having the knowledge represented by attributes from $B$; $\overline{B}X$ is the set of elements of $U$ which can be possibly classified as elements of $X$, employing the knowledge represented by attributes from $B$; set $BN_B$(X) is the set of elements which cannot be classified either to $X$ or to -$X$ having

knowledge *B*. If $X_1$ ,...,$X_{r(d)}$ are decision classes of A then the set, $\{\underline{B}X_1 \cup ... \cup \underline{B}X_{r(d)}\}$ is called the *B-positive region* of A and is denoted by $POS_B(d)$. If $C \subseteq A$ than the set $POS_B(C)$ is defined as $POS_B(d)$ where $d(x)=\{a(x):a \in C\}$ for $x \in U$ is an attribute representing the set *C* of attributes. If $B,C \in A$, then $B \rightarrow_{A,k} C$ where $k = \frac{|POS_B(C)|}{|U|}$ denotes the *partial dependence* of *C* on *B*. One can measure the importance of an attribute *a* with respect to the decision *d* in a given decision table as $|POS_a(d)| / |U|$. Vagueness of a set (category) is due to the existence of a boundary region. The following qualities of the lower approximation of *X* by *B* in A and upper approximation of *X* by *B* in A were introduced in:

$$\underline{\gamma}_B(X) = \frac{|\underline{B}X|}{|U|} \qquad \text{and} \qquad \overline{\gamma}_B(X) = \frac{|\overline{B}X|}{|U|}$$

Thus, the quality of lower approximation of *X* by *B* in A is the ratio of the number of all certainly classified objects by attributes from *B* as being in *X* to the number of all objects in the system. $\underline{\gamma}_B(X)$ is intended to capture the degree of completeness of our knowledge about the set *X*. It is a kind of relative frequency. The quality of upper approximation of *X* by *B* in A is the ratio of the number of all possibly classified objects by attributes from *B* as being in *X* to the number of all objects in the system. It is also a kind of relative frequency. One can also consider another measure of the set vagueness with respect to a given set *B* of attributes: $\alpha_B(X) = \frac{|\underline{B}X|}{|\overline{B}X|}$ , If A=($U$,$A \cup \{d\}$) is a decision table then we define a function $\delta_A(x):U \rightarrow P(\{1,...,r(d)\}$, called the generalized decision in A, by $\delta_A(x) = \{i: \exists x' \in U\ x'\ IND(A)\ x$ and $d(x)=i\}$. A decision table A is called *consistent* (*deterministic*) if $|\delta_A(x)=1|$ for any $x \in U$, otherwise A is *inconsistent* (*non-deterministic*). It is easy to see that a decision table A is consistent iff $POS_A(d)=U$. Moreover, if $\delta_B = \delta_{B'}$ then $POS_B(d)=POS_{B'}(d)$ for any non-empty sets $B,B' \subseteq A$. A subset *B* of the set *A* of attributes of decision table A = ($U$, $A \cup \{d\}$) is a relative reduct of A iff *B* is a minimal set with the following property : $\delta_B = \delta_A$. The set of all relative reducts in A is denoted by *RED*(A,*d*).

## 3. Expalainatory Example

Assume a dataset D viewed, as a table where attributes are columns and objects are rows, as in Table 1. U denotes the set of all objects in the dataset. A is the set of all attributes. C is the set of conditional (or input) attributes, and D is the set of decision attributes. U = {0, 1, 2, 3, 4, 5, 6, 7}, A = {a, b, c, d, e}, C = {a, b, c, d} and D = {e}. f(x, q) denotes the value of attribute q ∈ A in object x ∈ U. f(x, q) defines an equivalence relation over U. For instances:
$R_a$ = {{1, 7}, {0, 3, 4}, {2, 5, 6}}, $R_b$ = {{0, 2, 4}, {1, 3, 6, 7}, {5}}, $R_c$ = {{2, 3, 5}, {1, 6, 7}, {0, 4}}, $R_d$ = {{4, 7}, {1, 2, 5, 6}, {0, 3}}, and $R_e$ ={{0}, {2, 4, 5, 7}, {1, 3, 6}}. Assume a subset of the set of attributes, P ⊂ A. For instance, if P = {b, c}, objects 0 and 4 are indiscernible; 1, 6 and 7 likewise. The rest of the objects are not. This applies to the example dataset as follows:

| X∈U | a | b | c | d | e |
|---|---|---|---|---|---|
| 0 | 1 | 0 | B | B | K |
| 1 | 0 | 1 | 1 | 1 | Z |
| 2 | B | 0 | 0 | 1 | Y |
| 3 | 1 | 1 | 0 | B | Z |
| 4 | 1 | 0 | B | 0 | Y |
| 5 | B | B | 0 | 1 | Y |
| 6 | B | 1 | 1 | 1 | Z |
| 7 | 0 | 1 | 1 | 0 | Y |

Table 1

$U/IND(P) = U/IND(b) \otimes U/IND(c) =$
$= \{\{0, 2, 4\}, \{1, 3, 6, 7\}, \{5\}\} \otimes \{\{2, 3, 5\}, \{1, 6, 7\}, \{0, 4\}\} =$
$= \{\{0, 2, 4\} \text{ I } \{2, 3, 5\}, \{0, 2, 4\} \text{ I } \{1, 6, 7\}, \{0, 2, 4\} \text{ I } \{0, 4\}, \cdots \{5\} \text{ I } \{0, 4\}\} = \{\{2\}, \{0, 4\}, \{3\}, \{1, 6, 7\}, \{5\}\}$

*If* P = {a, b, c}, *then*, similarly:
$U/IND(P) = U/IND(a) \otimes U/IND(b) \otimes U/IND(c)$
Assuming P and Q are equivalence relations in U, the positive, negative and boundary regions are defined as ($POS_P$ (Q), $NEG_P$ (Q) and $BN_P$ (Q) respectively) as:

Fig. 1 illustrated these concepts. For example, assuming:
P = {b, c} and Q = {e}:
$POS_{IND(P)}(IND(Q)) = Y \{\{\}, \{2, 5\}, \{3\}\} = \{2, 3, 5\}$

$NEG_{IND(P)}(IND(Q)) = U - Y \{\{0, 4\}, \{2, 0, 4, 1, 6, 7, 5\}, \{3, 1, 6, 7\}\} = \{\}$.
$BN_{IND(P)}(IND(Q)) = U - \{2, 3, 5\} = \{0, 1, 3, 4, 6, 7\}$.
Also the degree of dependency of a set Q of decision attributes on a set of conditional attributes P defined with $\gamma_P$ (Q), the complement of $\gamma$ gives a measure of the contradictions in the selected subset of the dataset. If $\gamma = 0$, there is no dependence; for $0 < \gamma < 1$, there is a partial dependence. If $\gamma = 1$, there is complete dependence. For instance, in the example:

$$\gamma_{\{b,c\}}(\{e\}) = \frac{\| \{2,3,5\} \|}{\| \{0,1,2,3,4,5,6,7\} \|} = \frac{3}{8} = 0.375$$

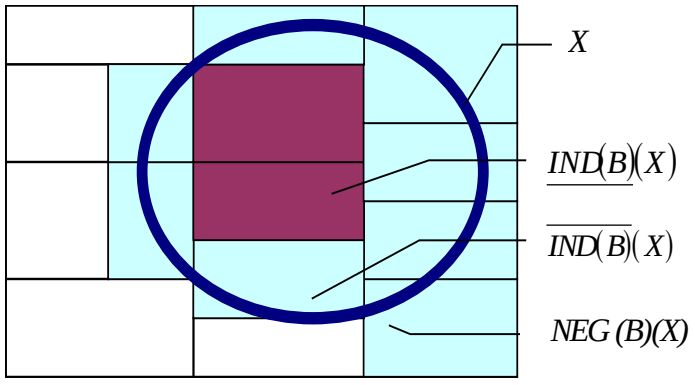


Fig. 2: The Lower And Upper Approximation Of Set X

This shows that, of the 8 objects, only 3 can be classified into the decision attribute e, given conditional attributes b and c. The other 5 objects (0,1,4,6,7) represent contradictory information. It is now possible to define the significance of an attribute. This is done, by calculating the change of dependency when removing the attribute from the set of considered conditional attributes. Given P, Q and an attribute $x \in P$:

$\sigma_P(Q,x) = \gamma_P(Q) - \gamma_{P-\{x\}}(Q)$. The higher the change in dependency, the more significant x is. For example, let P = {a, b, c} and Q = {e}. $\gamma_{\{a,b,c\}}(\{e\}) = 4/8$, $\gamma_{\{a,b\}}(\{e\}) = 4/8$, $\gamma_{\{b,c\}}(\{e\}) = 3/8$, $\gamma_{\{a,c\}}(\{e\}) = 4/8$. Using these calculations, it is possible to evaluate the significance of the three conditional attributes a, b and c as, $\sigma_P$ (Q, a) = 1/8, $\sigma_P$ (Q, b) = 0, $\sigma_P$ (Q, c) = 0, this shows that attribute a is not indispensable, having a significance of *0.125*, while attributes b and c can be dispensed with, as they do not provide any information significant for the classification into e. Attribute reduction involves removing attributes that have no significance to the classification at hand. It is obvious that a dataset may have more than one attribute reduct set. The set of reducts R is defined as: $R = \{X: X \subseteq C, \gamma_C(D) = \gamma_X(D)\}$. It is evident from this that the rough sets theory, will not compromise with a set of conditional attributes that has a large part of the information of the initial set, C - it will always attempt to

reduce the attribute set while losing no information significant to the classification at hand. To force the reduct to be the smallest possible set of conditional attributes, the minimal reduct. Rmin $\subseteq$ R is specified as the set of shortest reducts: Rmin = $\{X : X \in R, \forall Y \in R, \| X \| \leq Y\}$.

## 4. The Developed Algorithm

**Procedure** select the best structure of rule generator system for data mining problems.
**Input :**The Information System (decision table).
**Output :** Best rules for the data mining problems and accurate performance.
**Begin:**
**Process1:**
1.1 Input the decision table.
1.2 Reduct the data set (features selection): (its out put is reduced decision table).
**Process 2:**
2.1 Construct the rule generator according to the reduced decision table.
2.2 Test the constructed rule generator system.
2.3 **If** the total accuracy of the system is well and acceptable **then End**
**else goto** process 2.2.
**End.**

## 5. The Data Sets And Experimental Results

Experiments have been performed on data sets of the PROBEN1 benchmark set of real-world problems [5] originated from the UCI Machine Learning Repository [6]. Table 2. shows the different data sets used for the experiments. The first database, known as *cance*r, addresses a very important problem in the medical domain, the breast cancer diagnosis. The purpose is to find intelligible rules to classify a tumor as either benign or malignant. It is constituted by 699 examples of which 458 are benign and 241 are malignant examples. Each instance contains 10 integer-valued attributes of which the first is the diagnosis class, while the other nine attributes are related to cell descriptions gathered by microscopic examination. The second data set, designated as *glas*s, classifies glass types. It is constituted by 214 instances, 6 classes, and 9 attributes related to chemical analysis of glass sprinters plus the refractive index. Table 2. shows these data sets and their description. While Table 3. shows the experimental results for the attributes reductions, and the system accuracy with rule generation using See 5.

| Problem | *Attr* | *Con* | *Dis* | *Classes* | *Sampls* | *Training S* | *Test S* |
|---|---|---|---|---|---|---|---|
| *Cancer* | 9 | 0 | 9 | 2 | 699 | 525 | 174 |
| *Glass* | 9 | 9 | 0 | 6 | 214 | 161 | 53 |

Table 2: The Data Sets Used And Their Descriptions.

| Data set | # of features | Selected features | Rejected features | accuracy |
|---|---|---|---|---|
| Cancer | 10 | 5 | 4 | 93.1% |
| Glass | 9 | 4 | 5 | 92.4% |

Table 3: Experimental results for attributes reduction and systems accuracy using See 5.

## 6. Conclusion

From these results shown in Table 3, filtering attributes with rough set data analysis, which improves the strength of the results (i.e.the accuracy of the rules). It was found that we could make simple classifiers and less complex in both of time and space, using rough sets as data reduction method.

## Refrences:


1. Fayyad,U., Piatetsky-Shapiro,G. and Smyth, P. Knowledge Discovery and Data Mining: Toward a Unifying Framework. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, 1996, pp. 82-88
2. Z. Pawlak. Rough sets. International Journal of Computer and Information Sciences, 11(5): October 1982, pp. 341–356
3. Q. Shen, A. Chouchoulas. Combining rough sets and data-driven fuzzy learning. PatternRecognition, 32(12): 1999, pp. 2073–2076
4. Z. Pawlak. Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers
5. L. Prechelt, PROBEN1—a set of neural network benchmark 805 problems and benchmarking rules, Technical Report 21/94, 806 faculty information, germany, 807,1994
6. P. Murphy, D.W. Aha, UCI repository of machine learning databases [WWW page], http://www.ics.uci.edu/mlearn/MLRepository.html