Data Filtration:  
A Rough Set Approach

Andrzej Skowron  
Institute of Mathematics  
Warsaw University  
Banacha 2, 02-097 Warsaw, Poland  
e-mail: skowron@mimuw.edu.pl

Abstract  
We show how to apply some near-to-functional relations between data to data filtration. A method for searching for new classifiers (features) is described. It is based on searching for some functions approximating near-to-functional relations.

1 Introduction  
We propose a rough set approach to data filtration extending some ideas developed in such areas as signal filtration or image compression [14] and mathematical morphology [7], [15], [6]. We present a general strategy for filtering data in a given decision table on the basis of discovered local near-to-functional relations between data.

We discuss also a method for searching for classifiers in the set of approximation functions representing some near-to-functional relations between data. This method seems to be a promising tool for automatic extracting of classifiers from decision tables.

2 Rough Set Preliminaries  
Information systems [4] are used for representing knowledge. Rough sets have been introduced [4] as a tool to deal with inexact, uncertain or vague knowledge in artificial intelligence applications.

An information system is a pair \( \mathbb{A} = (U, A) \), where \( U \) is a non-empty, finite set of objects and \( A \) – 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 \( a \). By \( V \) we denote the set \( \bigcup \{ V_a : a \in A \} \).

Any information system \( \mathbb{A} = (U, A) \) and non-empty set \( B \subseteq A \) determine a \( B \)-information function

\[
Inf_B : U \to \mathbb{P}(B \times \bigcup_{a \in B} V_a)
\]

defined by \( Inf_B(x) = \{(a, a(x)) : a \in B\} \). We write \( Inf_B \) instead of \( Inf_B^A \) when no confusion arises. The set \( \{Inf_B(x) : x \in U\} \) is denoted by \( INF(\mathbb{A})|B \).
Elements of $INF(A)|B$ are called information vectors of $A$ restricted to $B$. The set $INF(A)|A$ will be denoted by $INF(A)$. By $INF(A,V)$ we denote the set of all functions $a$ from $U$ into $V$ satisfying $a(x) \in V_a$ for any $x \in U$ and $a \in A$.

We consider a special case of information systems called decision tables. A decision table is an information system of the form $A = (U, A \cup \{d\})$, where $d \notin A$ is a distinguished attribute called the decision. The elements of $A$ are called conditions.

The cardinality of the image $d(U) = \{k : d(s) = k \text{ for some } s \in U\}$ is denoted by $r(d)$. We assume that the set $V_d$ of values of the decision $d$ is equal to $\{I, \ldots, r(d)\}$.

Let us observe that the decision $d$ determines a partition $CLASS_A(d) = \{X_1, \ldots, X_{r(d)}\}$ of the universe $U$, where $X_k = \{x \in U : d(x) = k\}$ for $1 \leq k \leq r(d)$. The set $X_i$ is called the $i$-th decision class of $A$.

Let $A = (U, A)$ be an information system. With every subset of attributes $B \subset A$, an equivalence relation, denoted by $IND_A(B)$ (or $IND(B)$) called the $B$-indiscernibility relation, is associated and it is defined by

$IND(B) = \{(s, s') \in U^2 : \text{ for every } a \in B, a(s) = a(s')\}$

Objects $s, s'$ satisfying the relation $IND(B)$ are indiscernible by attributes from $B$.

If $A = (U, A)$ is an information system, $B \subset A$ is a set of attributes and $X \subset U$ 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 the $B$-lower and the $B$-upper approximation of $X$ in $A$, and they are denoted by $BX$ and $BX_\uparrow$, respectively.

The set $BN_B(X) = \overline{BX} - BX$ will be called the $B$-boundary of $X$. When $B = A$ we also write $BN_A(X)$ instead of $BN_A(X)$.

Sets which are unions of some classes of the indiscernibility relation $IND(B)$ are called definable by $B$ (or, $B$-definable, in short). A set $X$ is thus $B$-definable iff $\overline{BX} = BX$. Some subsets (categories) of objects in an information system cannot be exactly expressed by employing available attributes but they can be defined roughly.

If $A = (U, A \cup \{d\})$ is a decision table and $B \subset A$, then we define a function $\partial_B : U \rightarrow \mathcal{P}(\{1, \ldots, r(d)\})$, called the $B$-generalized decision in $A$, by

$\partial_B(x) = \{i : \exists x' \in U : x'IND(B)x \text{ and } d(x) = i\}$

The $A$-generalized decision $\partial_A$ in $A$ is called the generalized decision in $A$.

A decision table $A$ is called consistent (deterministic) if $|\partial_A(x)| = 1$ for any $x \in U$, otherwise $A$ is inconsistent (non-deterministic).

Now we recall the definition of decision rules. Let $A = (U, A \cup \{d\})$ be a decision table and let $V = \bigcup\{V_a : a \in A\} \cup V_d$.

The atomic formulas over $B \subset A \cup \{d\}$ and $V$ are expressions of the form $a = v$, called descriptors over $B$ and $V$, where $a \in B$ and $v \in V_a$. The set $\mathcal{F}(B, V)$ of formulas over $B$ and $V$ is the least set containing all atomic formulas over $B$ and $V$ and closed with respect to the classical propositional connectives $\lor$ (disjunction), $\land$ (conjunction), and $\neg$ (negation).