SELECTION OF OBJECTS AND ATTRIBUTES A TOLERANCE ROUGH SET APPROACH

Marek Krętowski and Jarosław Stepaniuk
Institute of Computer Science, Bialystok University of Technology
Wiejska 45A, 15-351 Bialystok, Poland
e-mail: {mkret, jstepan}@ii.pb.bialystok.pl

ABSTRACT

We present a method allowing to reduce the number of examples and the number of attributes involved in the process of learning from examples. This method is based on a generalization of the rough set approach. We introduce a tolerance binary relation at the level of values of any single attribute. Next at the level of the set of all (conditional) attributes, we define an operator aggregating the tolerance relations defined by single attributes. The tolerance relations and aggregating operators create parameters for our method which should be tuned to obtain a high quality of object classification. We discuss a method of searching for relevant tolerance relations on attribute values.

INTRODUCTION

In standard rough sets [5] introduced by Pawlak an equivalence relation on the universe of objects is defined based on their attributes values. In particular, this equivalence relation is constructed based on the equality relation on attribute values. Many attempts were made to resolve limitations of this approach and many authors proposed interesting extension of the initial model [2, 6, 9, 13, 11]. Generalization of rough sets which is discussed in this paper concerns tolerance relation between both attribute values and objects. Very important feature of each knowledge discovery system is how it treats different types of attribute, especially cardinal ones [10, 14]. Many of such systems use preliminary discretization (quantization). Our approach based on tolerance relation allows examine data without such pre-processing.

In real world databases amount of information is raising rapidly. Many algorithms of knowledge discovery have high complexity, so today's very fast computers are not able to process all possible data. Not only from this point of view development of efficient methods for data reduction is crucial for progress in knowledge discovery from large

experimental data sets. We propose a data reduction technique whose aim is to reduce the number of examples and the number of attributes involved in the process of learning from examples.

Let $A = (U, A \cup \{d\})$ be a decision table [5], where U is a set of objects (examples), A is a set of condition attributes and d is a decision. The reduction process of A consists of finding a new decision table $A' = (U', A' \cup \{d\})$ that satisfies the conditions $U' \subset U$, $A' \subset A$ and the decision rules constructed from A' have (almost) the same quality of classification as the decision rules constructed from A. The elements which belong to the new decision table are chosen using an evaluation criterion based on rough set theory [5] and Boolean reasoning. More precisely we use the notions of a tolerance attribute reduct [8] and an absorbent set of object set [12, 2].

This paper is organized as follows. In Section 1 basic notions concerning rough set concept based on tolerance relations are presented. Construction of tolerance relation is described in Section 2. In Section 3 we discuss the problem of searching for optimal tolerance relation and its efficient solution using genetic algorithm. In Section 4 we investigate some attribute reduction problems for tolerance decision tables. We also consider problem of the number of objects' reduction and we propose a procedure based on the notion of a relative absorbent set.

1. TOLERANCE ROUGH SETS

In this section we present basic notions of the rough set concept based on tolerance relations [8, 3]. The standard rough set model can be generalized by considering any type of binary relations on attribute values, instead of the trivial equality relation [8, 11, 13]. We propose a tolerance relation on attributes values in information system [2, 9] and similar approach to indiscernibility of objects in information system based on tolerance relation between them as more general extension of model described in [13].

Let $A=(U, A \cup \{d\})$ be a decision table, let V_a be a set of values of attributes of $a \in A$ and let $r(d)$ be a number of decision values.
Let $\mathcal{R}_A=\{R_a: R_a \subseteq V_a \times V_a \text{ and } a \in A\}$ be a set of tolerance relations. Each such relation is:

- reflexive (for all $v \in V_a$ vR_av)
- symmetric (for all $v, v' \in V_a$ if vR_av', then $v'R_av$).

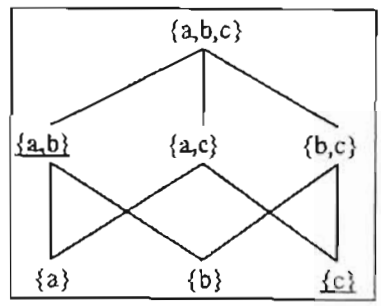
We say that two values $a(x)$ and $a(y)$ are „similar" $\Leftrightarrow a(x)R_a a(y)$.

Let Ψ_A be a family of subsets of A such that for all $C, C' \in \Psi$ if $C \neq C'$, then $C \not\subset C'$ and $C' \not\subset C$.
Global relations on the set of objects can be defined as follows ([8, 13]):

$x \tau(\mathcal{R}_A) y$ iff $\forall a \in A (a(x)R_a a(y))$

and more general

$x\tau(\mathcal{R}_A, \Psi_A)y$ iff $\exists C \in \Psi_A \forall a \in C(a(x)R_a a(y))$.



Let us observe that $\tau(\mathcal{R}_A)$ and $\tau(\mathcal{R}_A, \Psi_A)$ are tolerance relations. Sometimes we will denote $\tau(\mathcal{R}_A, \Psi_A)$ by τ_A., in short.

Example 1.1

Let for every attribute $a \in A$ a cost of measurement of value of a is defined. Let $cost : A \rightarrow N$, where N is the set of natural numbers. Let $cost(C)$ be the sum of all costs of attributes belonging to C. Let l be a given natural number. We define Ψ_A as follows $C \in \Psi_A$ iff $(cost(C) \leq l$ *and for all* $C' \neq C$ *and* $C \subset C'$, $cost(C') > l)$.

Let for example $cost(a)=1$, $cost(b)=2$, $cost(c)=3$, then for $l=4$ we obtain $\Psi_A=\{\{a,b\}, \{a,c\}\}$ and for $l=3$ $\Psi_A=\{\{a,b\}, \{c\}\}$. □

A tolerance decision table is defined by $(A, \tau(\mathcal{R}_A, \Psi_A))$, where $\tau(\mathcal{R}_A, \Psi_A)$ is a tolerance relation on the set of objects. We define tolerance set determined by an object x as follows:

$$TS(x) = \{x' \in U: x\tau(\mathcal{R}_A, \Psi_A)x'\}.$$

$TS(x)$ contains all objects similar to x. Tolerance generalized decision is defined as follows:

$$\partial_\tau(x) = \{i: \exists x' \in U \; x\tau(\mathcal{R}_A, \Psi_A)x' \; and \; d(x') = i\}.$$

Relative absorbent set

A subset $R_{abs} \subseteq U$ is a relative absorbent set for $(\tau(\mathcal{R}_A, \Psi_A), d)$ iff

1) for each $x \in U$ there exists $y \in R_{abs}$ such that $x\tau(\mathcal{R}_A, \Psi_A)y$ and $d(x)=d(y)$,

2) for every proper subset $R_{abs}' \subset R_{abs}$ condition 1) is not true.

In the paper we are interested in minimal (with respect to cardinality) relative absorbent sets.

The set approximations ([8])

The lower approximation of $Y \subseteq U$ by $\tau(\mathcal{R}_A, \Psi_A)$ is defined as follows:

$$\underline{\tau(\mathcal{R}_A, \Psi_A)}(Y) = \bigcup_{x \in U} \{TS(x): TS(x) \subseteq Y\}.$$

The upper approximation of $Y \subseteq U$ by $\tau(\mathcal{R}_A, \Psi_A)$:

$$\overline{\tau(\mathcal{R}_A, \Psi_A)}(Y) = \bigcup_{x \in U} \{TS(x): TS(x) \cap Y \neq \emptyset\}.$$

The set $\underline{\tau(\mathcal{R}_A, \Psi_A)}(Y)$ is the set of all elements of U which can be with certainty classified as elements of Y, with respect to $\tau(\mathcal{R}_A, \Psi_A)$. The set $\overline{\tau(\mathcal{R}_A, \Psi_A)}(Y)$ is the set of

elements of U which can be possibly classified as elements of Y, employing knowledge included in $\tau(\Re_A, \Psi_A)$.

Let $Y_i = \{x \in U: d(x)=i\}$. The set

$$POS(\tau(\Re_A, \Psi_A), \{d\}) = \bigcup_{x \in U} \{TS(x): \exists i\, TS(x) \subseteq Y_i\} = \bigcup_{i=1}^{r(d)} \underline{\tau(\Re_A, \Psi_A)}(Y_i)$$

is called *$\tau(\Re_A, \Psi_A)$ - positive region of partition $\{Y_i: i=1,...,r(d)\}$*. The coefficient

$$\gamma(\tau(\Re_A, \Psi_A), \{d\}) = \frac{card(POS(\tau(\Re_A, \Psi_A), \{d\}))}{card(U)}$$

is called the quality of approximation of classification *$\{Y_i: i=1,...,r(d)\}$*. It expresses the ratio of all *$\tau(\Re_A, \Psi_A)$ - correctly* classified objects to all objects in the table.

Relative tolerance reduct

A subset $R \in A$ is a relative tolerance reduct for $(\tau(\Re_A, \Psi_A), d)$ iff

1) $POS(\tau(\Re_A, \Psi_A), \{d\}) = POS(\tau(\Re_R, \Psi_R), \{d\})$,

2) for every proper subset $R' \subset R$ condition 1) is not true.

Proposition 1.2 For every tolerance relation $\tau(\Re_A, \Psi_A)$ the following conditions are true

a) $POS(\tau(\Re_A, \Psi_A), \{d\}) = \bigcup_{x \in U} \{TS(x): card(\partial_\tau(x)) = 1\}$.

b) If for all $x \in U$ $\quad card(\partial_\tau(x)) = 1$ then $\gamma(\tau(\Re_A, \Psi_A), \{d\}) = 1$.

c) $\gamma(\tau(\Re_A, \Psi_A)) \le \gamma\left(\tau\left(\{A\}, \left\{\{(v,v)\}: v \in V_a\right\}: a \in A\right)\right)$.

d) It is possible that $x\, \tau(\Re_A, \Psi_A)\, y$ and $\partial_\tau(x) \neq \partial_\tau(y)$.

2. CONSTRUCTIONS OF TOLERANCE RELATION

Construction of tolerance relation one can start from setting relations between attribute values for each attribute. We propose to use as a base similarity measures, which one can adopt to different types of given attributes. The similarity measures which are presented below are only examples. One can create new similarity measures depending on additional information about an attribute. Relationship between tolerance relation and similarity measures could be described as follows:

$$a(x) R_a a(y) \Leftrightarrow s_a(a(x), a(y)) \ge t(a)$$

where $a \in A$, $x,y \in U$, R_a - relation between attribute values of attribute a, $t(a) \in [0,1]$ - similarity threshold for values of attribute a.

Let $A=(U, A \cup \{d\})$ be a decision table and let $r(d)$ be a number of decision values. We define similarity measures between two values of a given attribute $a \in A$.

For attribute $a \in A$ with numeric values one can define a similarity measure

$$s_a(v_i, v_j) = 1 - \frac{|v_i - v_j|}{|a_{\max} - a_{\min}|},$$

where a_{min}, a_{max} denotes the minimum and maximum values of attribute a, respectively. Assuming that the values of attribute a are ordered as follows $v_1 \le v_2 \le \ldots \le v_{card(V_a)}$ we let

$$s_a(v_i, v_j) = 1 - \frac{|i-j|}{card(V_a) - 1}.$$

For attribute a with nominal (categorical) values we consider the following similarity measures

$$s_a(v_i, v_j) = 1 - \sum_{k=1}^{r(d)} \frac{\left| P(d=k, a=v_i) - P(d=k, a=v_j) \right|}{r(d) \cdot P(d=k)},$$

where $P(d=k, a=v_i)$ is a probability that value of attribute a is equal to v_i and decision is k, i.e. $P(d=k, a=v_i) = \frac{card(\{x \in U : d(x)=k \wedge a(x)=v_i\})}{card(U)}$

$$s_a(v_i, v_j) = 1 - \frac{\left| K(d|a=v_i) - K(d|a=v_j) \right|}{Ka_{\max} - Ka_{\min}} \text{ (see [1]),}$$

where $K(d|a=v_i) = \sum_{k=1}^{r(d)} P(d=k|a=v_i) \log_2 \frac{P(d=k|a=v_i)}{P(d=k)}$

and $Ka_{\max}, Ka_{\min}$ are the maximum and minimum K values as defined in formula of the attribute a.

We assume that two values v_i, v_j of attribute a are similar when $s_a(v_i, v_j) \ge t(a)$

To construct global tolerance relation (it means between objects) one should at first find out Ψ_A. Our idea is to concern such Ψ_A that the following condition is true:

$C \in \Psi_A$ iff $\left(\frac{card(C)}{card(A)} \le t \text{ and } \frac{card(C')}{card(A)} > t \text{ for every } C' \text{ such that } C' \supset C \text{ and } C' \ne C \right)$,

where $t \in [0,1]$ is a similarity threshold for objects. Then all subsets of Ψ_A have the same number of attributes depends on the value of t.

The Hamming distance between two objects $H_A(x,y)$ is the number of attributes from A where two objects have no similar values, i.e.

$$H_A(x,y) = card(\{a \in A : not(a(x) R_a a(y))\}) = card(\{a \in A : s_a(a(x), a(y)) < t(a)\}).$$

We define

$$s'_A(x,y)=\begin{cases}1 & if \quad \exists C\in\Psi_A \forall a\in C\left(a(x)R_a a(y)\right)\\ 0 & otherwise\end{cases}=\begin{cases}1 & if \quad 1-\dfrac{H_A(x,y)}{card(A)}\geq t\\ 0 & otherwise\end{cases}.$$

We say that x and y are not similar when not $x\tau(\mathcal{R}_A, \Psi_A)y$ or equivalently $s'_A(x,y)=0$. Choosing an appropriate similarity measure can be done by performing experiments with a given decision table.

3. SEARCHING FOR OPTIMAL TOLERANCE RELATION

In this section we present problem of finding the optimal tolerance relation and its effective solution based on genetic algorithms. The problem is formulated as follows:

Input:

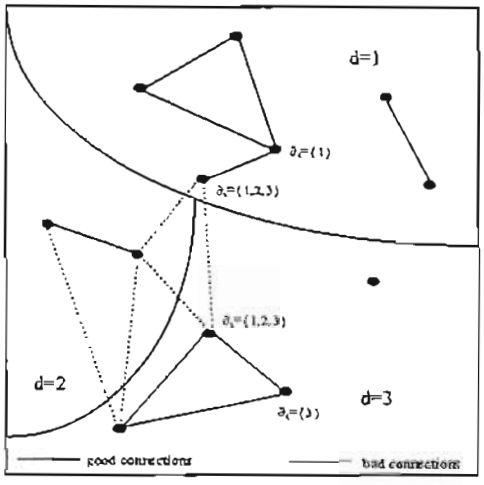
1) decision table $A=(U,A\cup\{d\})$

2) similarity measures $s_a: V_a\times V_a\rightarrow[0,1]$ for all $a\in A$.

Output: A set $\{t\}\cup\{t(a): a\in A\}$ of optimal thresholds. By optimal one can understand solution which satisfies different conditions. Actually we would like to obtain maximization of the following function:

$$\frac{card\left(\tau_A\cap\left\{(x,y): d(x)=d(y)\right\}\right)}{card\left(\left\{(x,y): d(x)=d(y)\right\}\right)}+\gamma\left(\tau_A,\{d\}\right).$$

First part of the objective function work for increasing the number of connections. But we are interested only in connections between objects with the same decision - good ones. Hence the second part of the function is introduced to prevent the positive region of partition. So the function tries to find out some kind of balance between enlarging τ_A and preventing $POS(\tau_A\{d\})$.



Before presenting solution of above problems we introduce some notions.

Connections. We use the notion of connections to express the indiscernibility of objects. We inherit it from very simple observation, that if $x\in TS(y)$ then $y\in TS(x)$ and then we can say that there is a connection between x and y. We propose to discern two kinds of connection between objects: „good" and „bad".

° x and y have good connection $\Leftrightarrow x\in TS(y)$ and $d(x)=d(y)$

° x and y have bad connection $\Leftrightarrow x \in TS(y)$ and $d(x) \neq d(y)$

Thresholds matrices. The thresholds matrix $TM(a)$ for $a \in A$ is $card(U) \times card(U)$ matrix, where $TM_{ij}(a) = s_a(a(x_i), a(x_j))$ - the highest value of threshold, which cause that x_i, x_j are indiscernible. Such matrix is symmetric ($TM_{ij}(a) = TM_{ji}(a)$) and what is more important number of different values in $TM(a)$ is less or equal $(k^2-k)/2+1$, where $k=card(V_a)$.

Tolerance sets matrices. Lets count tolerance sets (TS_a) separately for each $a \in A$. If we want to know $TS_a(x_i)$ we look for the i-th row in $TM(a)$ and if $TM_{ij}(a) \geq t(a)$ then $x_j \in TS_a(x_i)$. We can build $n \times n$ matrix for saving TS_a in such way:

$$TSM_{ij}(a) = \begin{cases} 1 \; if \; x_j \in TS_a(x_i) \\ 0 \; if \; x_j \notin TS_a(x_i) \end{cases}$$

Lets assume that we have chosen $t(a)$ for all $a \in A$ and using this thresholds we have built all TS_a. Then we can find general tolerance sets by building $TSM(A)$:

$$TSM_{ij}(A) = \begin{cases} 1 \; if \left(\sum_{a \in A} TSM_{ij}(a) \right) \geq t \cdot card(A) \\ 0 \text{ otherwise} \end{cases}$$

$TSM_{ij}(A)=1$ means that there is connection between x_i and x_j. This observation gives us a powerful tool to quick finding tolerance sets for given set of thresholds $\{t\} \cup \{t(a): a \in A\}$.

3.1. REDUCTION OF POSSIBLE THRESHOLD VALUES

If we increase the value of $t(a)$ then the TS_a will not change or become larger. So starting from $t(a)=1.0$ and decreasing the value of threshold we can using above property find all values when TS_a changes. We can create lists of such thresholds for each a. It is easy to obtain all possible threshold value for attribute a from $TM(a)$. We think that we can throw out some values of threshold and do not take them under consideration. Now we present a technique which help us to gain this aim.

Input: Descending lists of all values from $TM(a)$ for all $a \in A$ (1.0 is first value in each list).

Output: New lists of values with some values from initial one for all $a \in A$

Method: For each $a \in A$:

Step 1. For each value from list we examine what type of connections it introduce. We check what new connections appear when we decrease the threshold value from previous value in the list. It could be all „good" connections, all „bad" or „mixed" ones. So we can join with each value from list type of connections which this value introduce.

Step 2. Start from beginning of the list. Depending on connections type of value which is actually examined and its successor we decide if insert the value to new list or skip it. We use the following rule:

if actually examined value introduces only bad connections
or successor introduces only good ones
then skip the actual value
else insert the actual value into new list.

Step 3. To last value in the list which does not have successor we use rule as follows:

if type of connections ≠ bad
then insert the actual value into new list
else skip the actual value.

After executing above algorithm and if we set the main threshold t we can check all possible combinations of thresholds to find out the best for our purpose. Of course it will be long process, because in the worst case for only one t, the number of combinations is equal:

$$\frac{1}{2}\prod_{a\in A}\left(card(V_a)^2 - card(V_a)\right)+1.$$

So it shows that we need some heuristics to find, maybe not the best of all, but very good solution in reasonable time. We think that genetic algorithm will be suitable for this purpose.

3.2. GENETIC ALGORITHM FOR FINDING OPTIMAL TOLERANCE RELATION

We use standard schema of genetic algorithm (see [4]).

Representation. The individuals are represented by number strings of length *card(A)*. Each position in chromosome corresponds indirectly with value of threshold for attribute (in the i-th position there is a number from threshold list for i-th attribute). For example:

attribute	list of values	number of values in the list
a_1	- 1.0 0.9 0.87 0.6 0.4	5
a_2	- 0.95 0.5	2
a_3	- 1.0 0.98 0.95 0.93 ... 0.32	35

chromosome -	2	2	4
thresholds -	0.9	0.5	0.93
	$t(a_1)$	$t(a_2)$	$t(a_3)$

Initialization. For first population we used controlled random generator. It means that we accept only these individuals which have fitness greater or equal given threshold (for example ≥ 0.9).

Fitness function. The fitness function depends on two parameters:

- the number of good connections between objects,
- the quality of approximation of classification

$$Fitness(\tau_A) = \frac{card\left(\tau_A \cap \left\{(x,y): d(x)=d(y)\right\}\right)}{card\left(\left\{(x,y): d(x)=d(y)\right\}\right)} + \gamma\left(\tau_A, \{d\}\right).$$

Selection. We use modified tournament selection algorithm. It means that to select one chromosome we randomly choose k individuals from population (with equal probabilities) and then with probability P_t one with the best fitness wins or with probability $1\text{-}P_t$ we select any from k.

Mutation. Standard mutation affect with probability P_m of mutation on a single position of chromosome. Mutation of one position means replacement existing number by randomly chosen .

Crossing-over. We used classical, two-point crossover for chromosomes selected with the probability of P_c. For example:

$$\begin{array}{l} 3\ 5 \mid 0\ 23\ 3 \mid 4 \\ \quad\ \Uparrow\ \ \Downarrow \\ 5\ 1 \mid 1\ 3\ \ 7 \mid 4 \end{array} \quad \Rightarrow \quad \begin{array}{l} 3\ 5 \mid 1\ 3\ \ 7 \mid 4 \\ \\ 5\ 1 \mid 0\ 23\ 3 \mid 4 \end{array}$$

To find out the optimal thresholds we repeat genetic algorithm for a few possible values of t. All essential values of t we can find out from the set

$$\left\{1 - i \cdot \frac{1}{card(A)-1} : i = 0,\ldots, card(A)-1\right\}.$$

One can choose the most suitable solution from them depending on the value of fitness function.

For example for

$$t = 1 - \frac{1}{card(A)-1}$$

two objects x and y are similar iff there is at most one attribute $a \in A$ such that values $a(x)$ and $a(y)$ are not similar. Let us also observe that very low values of t are not interesting.

4. DATA REDUCTION

In this section we present methods of object set and attribute set reduction. First we present problem of data reduction.

Input:
1) decision table $A = (U, A \cup \{d\})$
2) similarity measures $s_a : V_a \times V_a \to [0,1]$ for all $a \in A$
3) a set $\{t\} \cup \{t(a) : a \in A\}$ of tolerance thresholds
Output: reduced decision table $A' = (U', A' \cup \{d\})$ such that, $U' \subset U$ is a relative absorbent set of $(U, A \cup \{d\})$ and $A' \subset A$ is a relative reduct of $(U', A \cup \{d\})$.

The ways how to find out relative absorbent sets and relative reducts are similar. In both cases first we build specific matrix (table). Next we make reduction of superfluous entries in such matrices. We set an entry to be empty if it is a superset of another non-empty entry. At the end of this process we obtain the set *COMP* of the so called components. From the set of components the described type of reducts or absorbent sets can be generated by applying Boolean reasoning [8]. We present heuristics for computing one reduct (absorbent set) of the considered type with the minimal number of attributes [2]. These heuristics can produce sets which are supersets of considered reducts but the heuristics are much more efficient than the general procedure. We describe shortly mentioned heuristics using reduct as an example.

First we introduce a notion of a minimal distinction. By *minimal distinction* (*md*, in short) we understand minimal set of attributes sufficient to discern between two objects. Let us observe that minimal component *comp* consists of minimal distinctions and *card(comp)* is equal or greater than *card(md)*[2]. We say that *md* is *indispensable* if there is a component composed out of only one *md*. We include all attributes from indispensable *md* to *R*. Then from *COMP* we eliminate all these components which have at least one *md* equal to *md* in *R*. It is important that the process of selecting attributes to *R* will be finished when the set *COMP* will be empty. We calculate for any *md* from *COMP*:

$c(md) = w_1 c_1(md) + w_2 c_2(md)$, where

$$c_1(md) = \left(\frac{card\,(md \cap R)}{card\,(md)} \right)^p$$

$$c_2(md) = \left(\frac{card\,\left(comp \in COMP : \exists_{\,md' \subset comp} \; md' \subset (R \cup md)\right)}{card\,(COMP)} \right)^q$$

for some natural numbers p and q, for example we can assume $p = q = 1$.

The first function is a "measure of extending" *R*. Because we want to minimize cardinality of *R* we are interested in finding *md* with the largest intersection with actual *R*. In this way we always add to *R* almost minimal number of new attributes. The second measure is used to examine our profit after adding attributes from *md* to *R*. We want to include to *R* the most frequent *md* in *COMP* and minimize *COMP* as much as it is possible. When $c_2(md)=1$ then after "adding this *md*" to *R* we will obtain „pseudo-reduct" i.e. it can be a superset of a reduct.

4.1. SELECTION OF OBJECTS

The main difference between finding out one relative absorbent set and one relative reduct is in the way in which we calculate and interpret components. In case of the relative absorbent set we do not build the discernibility matrix, but we replace it by a similar table containing for any object x_i all objects similar to x_i and with the same decision:

$ST(x_i)=\{x_j : s^l_A(x_i, x_j)=1 \text{ and } d(x_i)=d(x_j)\}$

After reduction we obtain components as essential entries in *ST*. For *COMP* we can apply the algorithm used to compute a reduct assuming *card(md)=1*. We add to the constructed relative absorbent set any object which is the most frequent in *COMP* and then eliminate from *COMP* all components having this object. This process terminates when *COMP* is empty.

4.2. SELECTION OF ATTRIBUTES

We consider a special case of relative reducts which can be constructed from discernibility matrix DM by adding some constraints on the reduction of attributes from the entries of DM in such a way that in any entry after reduction there is enough attributes to discern between corresponding objects [2]. In practice the process of reduction of attributes set is organized as follows. First we modify the discernibility matrix. We set an entry to be empty in DM for any entry corresponding to two non discernible objects.

$$DM_{i,j} = \begin{cases} \left\{a \in A : s_a\left(a(x_i), a(x_j)\right) < t(a)\right\} & \text{if} \quad s^l_A(x_i, x_j) = 0 \wedge d(x_i) \neq d(x_j) \\ \varnothing & \text{if} \quad s^l_A(x_i, x_j) = 1 \vee d(x_i) = d(x_j) \end{cases}$$

Next we apply heuristics presented at the beginning of the Section 4.

CONCLUSIONS

This paper has focused attention on data selection methods. We proposed a new technique which exploits tolerance rough set theory. The object reduction eliminates objects which are very close (with respect to the tolerance distance) to the remaining objects in the absorbent set. Attribute reduction is done without changing the classification quality. The decision rules can be generated from the reduced data table by applying Boolean reasoning methods developed in [7]. One can expect that the classification quality (of the unseen so far objects) obtained by tolerance decision rules [9] generated from the reduced decision table is very close to the quality of classification of the rules generated from the original decision table.

REFERENCES

[1] Hu X., Cercone N., Rough Sets Similarity-Based Learning from Databases, Proceedings of the First International Conference on Knowledge Discovery and Data Mining, Montreal, Canada, August 20-21 1995, pp. 162-167.

[2] Krętowski M., Polkowski L., Skowron A., Stepaniuk J., Data Reduction Based on Rough Set Theory, Proceedings of the International Workshop on Statistics, Machine Learning and Knowledge Discovery in Databases, Crete, Greece, April 28-29 1995.

[3] Marcus S., Tolerance Rough Sets, Cech Topologies, Learning Processes, Bull. of the Polish Academy of Sciences Technical Sciences, Vol. 42, No. 3, 1994, pp. 471-487.

[4] Michalewicz Z., Genetic Algorithms + Data Structures = Evolution Programs, Springer Verlag 1994.

[5] Pawlak Z., Rough Sets. Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht, 1991.

[6] Polkowski L., Skowron A., Zytkow J., Rough Foundations for Rough Sets, In: Soft Computing, T.Y.Lin, A.M.Waldinger (eds.), San Diego Simulation Councils, Inc., 1995, pp. 55-58.

[7] Skowron A., A Synthesis of Decision Rules: Applications of Discernibility Matrices, Proceedings of the Second International Workshop on Intelligent Information Systems, Augustow, Poland, June 7-11, 1993, pp. 30-46.

[8] Skowron A., Stepaniuk J., Generalized Approximation Spaces, In: Soft Computing, T.Y.Lin, A.M.Waldinger (eds.), San Diego Simulation Councils, Inc., 1995, pp. 18-21.

[9] Stepaniuk J., Krętowski M., Decision System Based on Tolerance Rough Sets, Proceedings of the Fourth International Workshop on Intelligent Information Systems, Augustow, Poland, June 5-9, 1995.

[10] Słowiński R., (ed.) Intelligent Decision Support - Handbook of Applications and Advances of the Rough Sets Theory. Dordrecht: Kluwer 1992.

[11] Słowiński R., Vanderpooten D., Similarity Relation as a Basis for Rough Approximations, Proceedings of the Second Annual Joint Conference on Information Sciences, Wrightsville Beach, N. Carolina, USA, September 28 - October 1, 1995, pp. 249-250.

[12] Tentush I., On Minimal Absorbent Sets for Some Types of Tolerance Relations, Bulletin of the Polish Academy of Sciences Technical Sciences Vol. 43, No. 1, 1995, pp. 79-88.

[13] Yao Y.Y., Wong S.K.M., Generalization of Rough Sets Using Relationships Between Attribute Values, Proceedings of the Second Annual Joint Conference on Information Sciences, Wrightsville Beach, N. Carolina, USA, September 28 - October 1, 1995, pp. 30 - 33.

[14] Ziarko W., (ed.) Rough Sets, Fuzzy Sets and Knowledge Discovery, Springer Verlag 1994.