Searching for Relational Patterns in Data

Sinh Hoa Nguyen and Andrzej Skowron

Institute Mathematics, Warsaw University
Warsaw, 02-097, Banacha Str. 2
emails: {hoa,skowron}@mimuw.edu.pl

Abstract. We consider several basic classes of tolerance relations among objects. These (global) relations are defined from some predefined similarity measures on values of attributes. A tolerance relation in a given class of tolerance relations is optimal with respect to a given decision table **A** if it contains only pairs of objects with the same decision and the number of such pairs contained in the relation is maximal among all relations from the class. We present a method for (sub-)optimal tolerance relation learning from data (decision table). The presented method is based on rough set approach. We show that for some basic families of tolerance relations this problem can be transformed to a relative geometrical problem in a real affine space. Hence geometrical computations are becoming useful tools for solving the problem of global tolerance relation construction. The complexity of considered problems can be evaluated by the complexity of the corresponding geometrical problems. We propose some efficient heuristics searching for an approximation of optimal tolerance relations in considered families of tolerance relations. The global tolerance relations can be treated as patterns in the cartesian product of the object set.We show how to apply the relational patterns (global tolerance relations) in clustering and classification of objects.

1 Introduction

In rough set theory [10] the notion of set approximation has been introduced by using equivalence relation defined on the set of objects. In some cases, it is necessary to generalize this notion by using tolerance relation (similarity relation) [4, 15].The tolerance relation can be defined in many different ways. Often tolerance relations are given by introducing some local similarity measures on values of attributes together with some rules of composing those local similarities into global ones.

One of the main problem of methodology for data mining is to develop methods for automatic pattern extraction from data. In our previous papers [7, 8] we have suggested to search for such patterns in the form of templates. Using them it was possible to decompose a given table into a family of subtables corresponding to these patterns and to create subdomains of a given space of objects. The objects from any subdomain have many common features what suggests that they can create a "regular" subdomain for which strong decision rules can be generated.

In this paper we consider patterns defined by tolerance relations. These patterns correspond to some (sub-)optimal tolerance relations extracted from data. In this way we propose rather to search for (sub-)optimal tolerance relations from data in predefined classes of tolerance relations than by assuming apriori their form (as it is often done when clustering methods are used).

In searching for tolerance relations from data we follow a method proposed in [14] based on rough sets. We propose a method of searching for (sub-)optimal tolerance relation (with respect to the number of the pairs of objects with the same decision from this relation) by transforming the considered problem to the problem of approximate description of some regions in affine space R^k, where k is equal to the number of (conditional) attributes.

We consider several classes of tolerance relation. Any class is characterized by a first order formula and some parameters which are tuned in the optimization process. For any of these classes we propose strategies searching for (sub-)optimal tolerance relation in it i.e. described by a maximal set of object pairs having the same decision. We illustrate how the extracted patterns can be used for cluster construction and classification of new objects.

2 Basic notions

2.1 Rough set preliminaries

An *information system* is defined by a pair $\mathbf{A} = (U, A)$, where U is a non-empty, finite set of *objects* called *universe*, $A = \{a_1, \ldots, a_k\}$ is a non-empty, finite set of *attributes*, i.e. $a_i : U \to V_{a_i}$ for $i \in \{1, ..., k\}$, where V_{a_i} is called *the domain of the attribute* a_i.

The *information space* of A is defined by $INF_A = \prod_{a \in A} V_a$. We define the information function $Inf_A : U \to INF_A$ by

$$Inf_A(x) = (a_1(x), \ldots, a_k(x)), \text{ for any } x \in U.$$

Any object $x \in U$ is represented by its *information vector* $Inf_A(x)$.

A *decision table* $\mathbf{A} = (U, A \cup \{d\})$, where $d \notin A$ is a distinguished attribute called *decision* is a special case of information systems.

For any information system $\mathbf{A} = (U, A)$ and a subset $B \subseteq A$ the *B-indiscernibility relation* IND_B can be defined by

$$xIND_B y \Leftrightarrow \forall_{a \in B}[a(x) = a(y)]$$

Obviously, IND_B is an equivalence relation. We denote by $[x]_{IND_B} = \{y : \langle x, y \rangle \in IND_B\}$ the equivalence class defined by the object $x \in U$.

For any $X \subseteq U$ one can define the lower approximation and the upper approximation of X by

$$\underline{X} = \{x \in U : [x]_{IND_B} \subseteq X\};\ \overline{X} = \{x \in U : [x]_{IND_B} \bigcap X \neq \emptyset\}, \text{respectively.}$$

The pair $(\underline{X}, \overline{X})$ is referred to as the rough sets of X.

2.2 Tolerance relation

Indiscernibility relation is a useful tool of rough set theory, but in many cases it is not sufficient, in particular, when we deal with real value attributes. In this case almost every object differs from another. The equivalence classes divide universe into tiny classes. Therefore the description of a subset of U is complicated (with respect to the number of equivalence classes) and not enough general. The standard rough set model can be generalized by assuming any type of binary relation (on attribute values) instead of the equivalence relation.

In this paper we consider a family of relations $\tau \subseteq U \times U$ which are *reflexive* (i.e. $\forall_{x \in U} \langle x, x \rangle \in \tau$) and*symmetric* (i.e. $\forall_{x,y \in U}(\langle x, y \rangle \in \tau \Rightarrow \langle y, x \rangle \in \tau)$). Such relations are called the *tolerance relations.*

For any $x \in U$ the *tolerance class* $[x]_\tau$ can be defined by

$$[x]_\tau = \{y \in U : \langle x, y \rangle \in \tau\}$$

We say, that the tolerance relation τ *identifies* objects x and y if $\langle x, y \rangle \in \tau$; otherwise we say that it *discerns* them.

One can define the lower approximation and the upper approximation of any subset $X \subseteq U$ by

$$\underline{\tau(X)} = \{x \in U : [x]_\tau \subseteq X\}; \quad \overline{\tau(X)} = \{x \in U : [x]_\tau \cap X \neq \emptyset\}, \text{ respectively.}$$

Let $\mathbf{A} = (A, U \cup \{d\})$. be a decision table. A *similarity measure* for an attribute $a \in A$ is a positive function $\delta_a : U \times U \to \Re^+ \cup \{0\}$ satisfing the following conditions:

1. $\delta_a(x, x) = 0$;
2. $\delta_a(x, y) = \delta_a(y, x)$;

Having a family of similarity measures $\{\delta_{a_i}\}_{a_i \in A}$ the *tolerance relation* $\tau \subseteq U \times U$ can be defined by

$$\langle x, y \rangle \in \tau \Leftrightarrow \Psi_R(\delta_{a_1}(x, y), \delta_{a_2}(x, y), \ldots, \delta_{a_k}(x, y)) = \mathbf{true} \tag{1}$$

where $\Psi(\xi_1, \xi_2, ..., \xi_k)$ is a first order logic propositional formula and Ψ_R is its realization in a relational structure of real numbers such that $\Psi_R(0, ..., 0) = \mathbf{true}$.

By C_k we denote the set $\{(r_1, r_2, ..., r_k) \in R^k : 0 \leq r_i, \text{ for } i = 1, ..., k\}$. For any relation τ defined by (1), the *interpretation of* τ is defined by

$$\overline{\tau} := \{(r_1, r_2, ..., r_k) \in C_k : \Psi_R(r_1, r_2, ..., r_k) = \mathbf{true}\} \subseteq C_k. \tag{2}$$

One can define different tolerance relations using different formulas $\Psi(\xi_1, ..., \xi_k)$. We list some basic families of parameterized tolerance relations used in the paper:

1. $\langle x, y \rangle \in \tau(\varepsilon_1, ..., \varepsilon_k) \Leftrightarrow \forall_{a_i \in A}[\delta_{a_i}(x, y) \leq \varepsilon_i]$
2. $\langle x, y \rangle \in \tau(w_1, ..., w_k, w) \Leftrightarrow \sum_{a_i \in A} w_i \cdot \delta_{a_i}(x, y) + w \leq 0$
3. $\langle x, y \rangle \in \tau(w_1, ..., w_k, w) \Leftrightarrow \sum_{a_i \in A} w_i \cdot \delta^2_{a_i}(x, y) + w \leq 0$
4. $\langle x, y \rangle \in \tau(\varepsilon_1, ..., \varepsilon_k) \Leftrightarrow \exists_{a_i \in A}[\delta_{a_i}(x, y) \leq \varepsilon_i]$
5. $\langle x, y \rangle \in \tau(w) \Leftrightarrow \prod_{a_i \in A} \delta_{a_i}(x, y) \leq w$

where δ_{a_i} is a predefined similarity measure for $i = 1, ..., k$ and $\varepsilon_i, \varepsilon, w_i, w$ are real numbers, called *parameters.*

A tolerance relation $\tau \subseteq U \times U$ is *consistent* with a decision table $\mathbf{A} = (A, U \cup \{d\})$ if

$$\langle x, y \rangle \in \tau \Rightarrow (d(x) = d(y)) \vee (\langle x, y \rangle \in IND_A)$$

for any objects $x, y \in U$.

The relation τ is *optimal* in the family $\mathcal{T}$ for a given $\mathbf{A}$ if τ contains the *maximal number* of pairs of objects among tolerance relations from $\mathcal{T}$ consistent with $\mathbf{A}$.

3 Extraction of global tolerance relation from data

Let $\mathbf{A} = (A, U \cup \{d\})$ be a decision table and let δ_a be a similarity measure for any attribute $a \in A$. The problem of extracting a tolerance relation in a given class $\mathcal{T}$ is a searching problem for parameters such that the tolerance relation with the parameters found in searching process is *optimal.* The searching problem for the optimal tolerance relation is of high complexity.

Our goal is to search for a sub-optimal tolerance relation that *discerns* between all pairs of objects with different decisions and *identifies* maximal number of pairs of objects with the same decision.

In the first stage of tolerance relation construction, we define a new decision table $\mathbf{B}$ called the *similarity table.* The table $\mathbf{B} = (U', A' \cup \{D\})$ is defined assuming given $\mathbf{A}$ and the set of similarity measures $\{\delta_a\}_{a \in A}$ by

$$U' = U \times U; A' = \{\delta_a\}_{a \in A}; \text{ and } D(x, y) = \begin{cases} 0 \text{ if } d(x) = d(y) \\ 1 \text{ otherwise} \end{cases}$$

The set of objects of the similarity table $\mathbf{B}$ (for a table $\mathbf{A}$) is equal to the set of all pairs of objects from table $\mathbf{A}$ and the attribute values are the values of the similarity measure functions for pairs of objects. The new table has a binary decision. The decision value for any pair of objects is equal to 0 if the objects have the same decision in the original table $\mathbf{A}$, and 1 otherwise.

The searching problem for a *sub-optimal* tolerance relation for table $\mathbf{A}$ among relations from a given class $\mathcal{T}$ of tolerance relations can be considered as the problem of decision rule extraction from the decision table $\mathbf{B}$. We are looking for decision rules describing the decision class corresponding to $D = 0$, i.e. the class associated with pairs of objects of the table $\mathbf{A}$ with the same decision. Our goal is to search for the rule of the form $\Psi\left(a_1'(u), a_2'(u), \ldots, a_k'(u)\right) \Rightarrow (D(u) = 0)$ satisfied by as many as possible objects $u \in U'$.

Vit.A	Vit.C	Fruit	Vit.A	Vit.C	Fruit
1.0	0.6	Apple	2.0	0.7	Pear
1.75	0.4	Apple	2.0	1.1	Pear
1.3	0.1	Apple	1.9	0.95	Pear
0.8	0.2	Apple	2.0	0.95	Pear
1.1	0.7	Apple	2.3	1.2	Pear
1.3	0.6	Apple	2.5	1.15	Pear
0.9	0.5	Apple	2.7	1.0	Pear
1.6	0.6	Apple	2.9	1.1	Pear
1.4	0.15	Apple	2.8	0.9	Pear
1.0	0.1	Apple	3.0	1.05	Pear

Table 1. Apples and pears

4 Geometrical interpretation

In this section we show that some families of tolerance relations have clear geometrical interpretations, i.e. they can be described in a straightforward way as subsets of real affine space R^k. Therefore the searching problem for a sub-optimal tolerance relation can be reduced to searching for an approximate description of the corresponding subset of real affine space R^k.

For a decision table $\mathbf{A} = (U, A \cup \{d\})$ with k conditional attributes and a set $\{\delta_a\}_{a \in A}$ of predefined similarity measures we build the similarity table $\mathbf{B} = (U', A' \cup \{D\})$. Every object $u \in U'$ can be represented by a point $p(u) = \left[a_1'(u), ..., a_k'(u)\right] \in R^k$ of one of two categories "white" or "black". A point $p(u) \in R^k$ is "white" iff $\left\{u' \in U' : p(u') = p(u)\right\}$ is non-empty and it consists of objects with the decision $D = 0$ only; otherwise $p(u)$ is "black". Below we present a geometrical interpretations of some standard tolerance relations. As a similarity measures we take the functions: $\delta_a(x, y) = |a(x) - a(y)|$ for any attribute $a \in A$. We take as an example a table with two attributes representing the quantity of vitamin A and C in apples and pears.

We want to extract the similarities of fruits of one category. The data about apples and pears are shown in Figure 1.

Below we present a geometrical interpretations of some standard tolerance relations in the space of pairs of objects from the fruit table.

1. The relation, called the *descriptor conjunction,* is defined by

$$\langle x, y \rangle \in \tau_1(\varepsilon_{1,...,\varepsilon_k}) \Leftrightarrow \bigwedge_{a_i \in A} [\delta_{a_i}(x, y) \leq \varepsilon_i] \tag{3}$$

where $\varepsilon_1, ..., \varepsilon_k \in R^+$. The interpretation of $\tau_1(\varepsilon_{1,...,\varepsilon_k})$ (see (2)) is given by

$$\overline{\tau_1(\varepsilon_{1,...,\varepsilon_k})} = \{(r_1, ..., r_k) \in C_k : 0 \leq r_i \leq \varepsilon_i \text{ for } i = 1, ..., k\}$$

$\overline{\tau_1(\varepsilon_1, ..., \varepsilon_k)}$ is an **interval** in R^k with boundaries $\varepsilon_1, \varepsilon_2, ..., \varepsilon_k$; it is attached to the origin O of axes (Figure 2b). When $\varepsilon_1 = ... = \varepsilon_k = \varepsilon$ instead of general interval in R^k we have hypercubes (Figure 2a). By $\mathcal{T}_1$ we denote the family of all hypercubes and by $\mathcal{T}_2$ we denote the family of all relations defined by (3).

2. The *linear combination* relation is defined by

$$\langle x, y\rangle \in \tau_2(w_{1,...,}w_k, w) \Leftrightarrow \textstyle\sum_{a_i \in A} w_i \cdot \delta_{a_i}(x, y) + w \leq 0$$

where $w_1, ..., w_k, w \in R$. The interpretation of $\tau_2(w_1, ..., w_k, w)$ is given by

$$\overline{\tau_2(w_1, ..., w_k, w)} = \left\{(r_1, ..., r_k) \in C_k : \textstyle\sum_{i=1}^{k} w_i \cdot r_i + w \leq 0\right\}$$

Hence $\overline{\tau_2(w_{1,...,}w_k, w)}$ is a region below the hyperplane $\sum_{i=1}^{k} w_i \cdot x_i + w = 0$ in C_k (Figure 3a). By $\mathcal{T}_3$ we denote the family of all tolerance relations of the form $\tau_2(w_{1,...,}w_k, w)$.

3. A linear combination can be extended to the higher order combination. We consider the relation defined by *square combination of similarity measures*

$$\langle x, y\rangle \in \tau_3(w_{1,...,}w_k, w) \Leftrightarrow \sum_{a_i \in A} w_i \cdot \delta_{a_i}^2(x, y) + w \leq 0. \tag{4}$$

where $w_1, ..., w_k, w \in R$. The interpretation of $\tau_3(w_1, ..., w_k, w)$ is given by

$$\overline{\tau_3(w_1, ..., w_k, w)} = \left\{(r_1, ..., r_k) \in C_k : \textstyle\sum_{i=1}^{k} w_i \cdot r_i^2 + w \leq 0\right\}$$

Hence $\overline{\tau_3(w_1, ..., w_k, w)}$ is a region in C_k bounded by **ellipsoid** (Figure 3b). By $\mathcal{T}_4$ we denote the family of all tolerance relations of the form (4)

4. The next relation called "*min*" is defined by the formula

$$\langle x, y\rangle \in \tau_4(\varepsilon) \Leftrightarrow \min_{a_i \in A}\{\delta_{a_i}(x, y)\} \leq \varepsilon,$$

where ε is a non-negative real. Hence $\overline{\tau_4(\varepsilon)} = \bigcup_{i=1}^{k}\{(r_1, ..., r_k) \in C_k : r_i \leq \varepsilon\}$ is a **sum of bands** with boundaries $x_i = 0$ and $x_i = \varepsilon$ for $i = 1, ..., k$. By $\mathcal{T}_5$ we denote the family of all tolerance relations of the form $\tau_4(\varepsilon)$.

5. The tolerance relation τ_5 is defined by a disjunction of atomic formulas

$$\langle x, y\rangle \in \tau_5(\varepsilon_1, ..., \varepsilon_k) \Leftrightarrow \bigvee_{a_i \in A}[\delta_{a_i}(x, y) \leq \varepsilon_i],$$

where $\varepsilon_{1,...,}\varepsilon_k$ are non-negative real numbers. This relation is a generalization of the relation "*min*" (Figure 4a).

6. Our last example is a tolerance relation defined by

$$\langle x, y\rangle \in \tau_6(w) \Leftrightarrow \prod_{a_i \in A} \delta_{a_i}(x, y) \leq w,$$

where $w \in R^+$. The set $\overline{\tau_6(w)}$ is equal to $\{(r_1, ..., r_k) \in C_k : r_1 \cdot ... \cdot r_k \leq w\}$. Hence it is a region in C_k bounded by **hyperboloid** (Figure 4b).

5 Heuristics

The time complexity of the searching problem for optimal tolerance relation parametrized by k parameters for a set of n objects is $O(n^k)$, because we have to test all possible values of parameter vector, where the number of possible values for one parameter is usually $O(n)$. This time is not feasible, when the dimension of the problem is large (the number of points n and the dimension k of the space are large). We show, that the approximations of some tolerance relations can be constructed if its geometrical description is known. Below we present heuristics for two important tolerance relation classes.

5.1 Searching for description conjunction

The first example of tolerance relation classes is the descriptor conjunction $\tau(\varepsilon_1, ..., \varepsilon_k)$ (see (3)) having the following interpretation

$$\overline{\tau(\varepsilon_{1,...,}\varepsilon_k)} = \{(r_1, ..., r_k) \in C^k : 0 \leq r_i \leq \varepsilon_i \text{ for } i = 1, ..., k\} \tag{5}$$

One can see that for given $\varepsilon_{1,...,}\varepsilon_k$, the set (5) is included in the interval $I(\varepsilon_{1,...,}\varepsilon_k)$ from R^k. Our goal is to search for parameters $\varepsilon_{1,...,}\varepsilon_k$ such that the interval $I(\varepsilon_{1,...,}\varepsilon_k)$ consists of "white" points only and, at the same time, as many as possible of them.

We start from the empty interval I_0(with one null boundary, let $\varepsilon_1 = 0$). The idea of the algorithm is based on construction of a sequence of intervals by transforming any successive interval I_i to a new interval I_{i+1}. Among generated intervals, we choose the best one. Transformation is performed by gradual augmenting the parameter ε_1, and by decreasing some of the remaining parameters so, that the interval $I_i(\varepsilon_1, ..., \varepsilon_k)$ is still consisting of while points only and including as many as possible points. The algorithm can be presented as follows:

Input: The set of labeled points from the space R^k.
Output: Parameters $\{\varepsilon_1, ..., \varepsilon_k\}$ of the sub-optimal interval.

1. Set $\varepsilon_1 = 0$ and $\varepsilon_i = \infty$ for $i = 2, ..., k$.
2. Gradually augment the value of $\varepsilon_1$to obtain a new interval $I(\varepsilon_1, ..., \varepsilon_k)$
3. If the interval $I(\varepsilon_1, ..., \varepsilon_k)$ contains black points then decrease some of the remaining parameters to eliminate "black" points from interval $I(\varepsilon_1, ..., \varepsilon_k)$. We choose such parameters to optimize the number of "white" points belonging to the modified interval.
4. Repeat Step 2 until all possible values of ε_1 are checked. Return parameters $\varepsilon_1, ..., \varepsilon_k$ of the optimal interval among considered in Step 3.

The algorithm can be implemented in $O(n^2 \cdot k)$ time by using sorted lists of possible values of considered parameters.

5.2 Searching for linear combination (hyperplane)

Let us consider a *linear combination* tolerance relation $\tau(w_1, ..., w_k, w)$ ant its interpretation

$$\overline{\tau(w_{1,...,}w_k, w)} = \left\{(r_1, ..., r_k) \in C_k : \sum_{i=1}^{k} w_i \cdot r_i + w \leq 0\right\} \tag{6}$$

For given parameters $w_1, ..., w_k, w$ the formula (6) describes the set of points with positive coordinates lying below the hyperplane $H : \sum_{i=1}^{k} w_i \cdot x_i + w = 0$. This hyperplane is determined by $(k+1)$ parameters. Any hyperplane divides the space into two half-spaces. We say the hyperplane H is *satisfactory* if the half-space below the hyperplane H contains only "white" points. Our goal is to search for a satisfactory hyperplane H with the (semi-)maximal number of "white" points below it.

The algorithm starts with randomly chosen hyperplane $H = \sum_{i=1}^{k} w_i \cdot x_i + w$. We generate a set of satisfactory hyperplanes starting from H. Among them we choose the best one. The satisfactory hyperplane can be computed on two stages. At first we rotate the hyperplane H to obtain a new hyperplane (i.e. determines a new partition of point set). Then we translate it until the points below hyperplane are all "white". The hyperplane H can be rotated by fixing k parameters, for example $w, w_1, ..., w_{j-1}, w_{j+1}, ..., w_k$ and modifying only one parameter w_j. We are interested in a values of w_j such that the modified hyperplane determines a new partition of set of objects. The idea is based on observation that a point $p = [x_1, x_2, ..., x_k] \in C^k$ is below H iff $H(p) \leq 0$ i.e. $w_j \leq \frac{-\sum_{i \neq j} w_i \cdot x_i - w}{x_j}$. Let $w_j(p) = \frac{-\sum_{i \neq j} w_i \cdot x_i - w}{x_j}$ and

$$S = \left\{w_j(p) : p = \left[a_1^{'}(u), ..., a_k^{'}(u)\right] \text{ for any } u \in U^{'}\right\} \tag{7}$$

then any value w_j chosen from S determines a new hyperplane defining a new partition of the point set. The algorithm can be presented briefly as follows:

Input: The set of labeled points of the space R^k.
Output: Parameters $\{w, w_1, ..., w_k\}$ of the optimal satisfactory hyperplane.

1. $H(w, w_{1,...,}w_k)$:= randomly chosen hyperplane;
2. **for** (any $j = 1...k$)
   begin
   Construct the set S defined in (7) and sort S in increasing order;
   for (any positive $v \in S$)
   begin
   $w_j := v$;
   Translate $H(w, w_{1,...,}w_k)$ to a *good* position i.e. with all "white" points below it and calculate the number of those white points. The *fitness* of the hyperplane is equal to this number.
   end
   end

3. Among *good* hyperplanes we choose a hyperplane with maximal fitness.

The algorithm repeats the Loop 2 $O(k)$ times and every loop takes $O(k \cdot n)$ times. Therefore the complexity of proposed algorithm is $O\left(k^2 \cdot n\right)$, where n is a number of points and k is the dimension of a space.

6 Tolerance relation in classification problems

6.1 Clustering method

Let $\mathbf{A} = (U, A)$ be an information system. Given a *consistent* tolerance relation τ defined on the *universe* U we define its τ^* by $\tau^* = \bigcup_{n \geq 0} \tau^n$. The cluster C can be defined as the object set such, that **if** $x, y \in C$ **then** $x \ \tau^* \ y$. The clusters of the universe U can be constructed in a straightforward way

repeat
  Choose $x \in U$, $C = [x]_{\tau^*}$;
  $U = U \setminus C$;
until $U = \emptyset$

One can see that clusters determined by the algorithm are disjoint and they contain the objects with the same decision. We can use those clusters for classification of new cases in different ways. One example of classification strategy is presented below:

Step 1 : Every cluster C_i is characterized by its *center* $\mathbf{c}_i$ and its *mass* $\mathbf{m}_i$ (Number of objects belonging to the cluster C_i);
Step 2 : Define the distance function d;
Step 3 : For a new object x, the number $p_i(x) = \frac{m_i}{d(c_i, x)}$ is a gravitation power measure of the cluster C_i influencing the new object x. The new case x is classified to the cluster with the maximal gravitation power $p_i(x)$.

6.2 Classification: Nearest Neighbours Method

For a given tolerance τ and any object x one can define the set of neighbours of x in the tolerance sense. The set of neighbors of x is defined gradually as follows:
$NN_1(x) = \{y : y \tau x\}$
$NN_k(x) = \left\{y : \bigvee_{z \in NN_{k-1}(x)} x \tau z \wedge z \tau y\right\}$
Having a set of neighbours of the object x, one can classify x using different strategies, for example one can take a majority rule as the standard criterion. Classification process of new objects is presented below

Step 1 : Construct the set of neighbours $NN_k(x)$ of x for some k. We choose the value k in such way that the set $NN_k(x)$ contains no less then M objects from training set.

Step 2 : Use M nearest neighbours of x to vote for the decision on x. The object x is classified to the decision class supported by the maximal number of objects from the $NN_k(x)$.

7 Conclusion

We have presented a new approach for extraction relational patterns in data. These patterns are described by tolerance relations extracted from data. The searching problem for optimal patterns can be transformed to some geometrical problems because almost all standard tolerance relations can be described by some regions in the space R^k. Hence one can extract tolerance relations from data by constructing approximation of corresponding regions. We have proposed some heuristic for the some important parametrized relation classes. We are working on the implementation of proposed methods.

Acknowledgement: This paper was supported by the State Committee for Scientific Research grant, KBN 8T11C01011.

References

1. Bezdek J.C., Chuah S., Leep D. Generalized k-Nearest Neighbour Rule, Fuzzy Sets and Systems, 1,8(3), 1986, pp.237-256.
2. Cover T., Hart P., Nearest Neighbour Pattern Classification, IEEE Trans. Inf. Theory, Vol.13, 1967, pp. 21-27.
3. Hu X., Cercone N., Rough Set Similarity Based Learning from Databases. *Proc of The Fourth International Workshop on Rough Set, Fuzzy Set and Machine Discovery.* August 20-21,1995 Montreal, Canada, pp.162-167.
4. Krawiec K., Slowiński R., Vanderpooten D. Construction of Rough Classifires Based on Application of a Similarity Relation. *Proc. of The Fourth International Workshop on Rough Set, Fuzzy Set and Machine Discovery.* November 6-8,1996, Tokio, Japan, pp.23-30.
5. Lin T.Y. Neighbourhood system and approximation in database and knowled base systems, Proc. of The Fourth International Symposium on Methodologies of Intelligent System, 1989.
6. Marcus S., Tolerance Rough Sets, Cech Topologies, Learning Process, Bull. of The Polish Academy of Technical Sciences, Vol. 42, No. 3, 1994, pp.471-487.
7. Nguyen S. H., Nguyen T. T., Skowron A., Synak P., Knowledge Discovery by Rough Set Methods, *Proc. of The International Conference On Information Systems Analysis and Synthesis,* July, 22-26, 1996, Orlando, USA, pp.26-33.
8. Nguyen S. H., Polkowski L., Skowron A., Synak P., Wróblewski J., Searching for Approximate Description of Decision Classes, *Proc.of The Fourth International Workshop on Rough Sets,Fuzzy Sets,and Machine Discovery,* November 6-8, 1996, Tokyo, Japan, pp.153-161.
9. Pawlak Z. Rough Classification, International Journal of Man-Machine Studies, 20, pp. 469-483, 1984.
10. Pawlak Z., Rough Sets. Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht, 1991.

11. Polkowski L., Skowron A., Zytkow J., Tolerance Based Rough Sets, In: *Soft Computing*, T.Y.Lin, A.M. Wildberger (eds.), San Diego, Simulation Council, Inc., 1995, pp.55-58.
12. Stepaniuk J., Similarity Based Rough Sets and Learning, *Proc. of The Fourth International Workshop on Rough Sets, Fuzzy Sets and Machine Discovery*, November 6-8, 1996, Tokyo, Japan, pp.18-22.
13. Stepaniuk J., Krętowski M., Polkowski L., Skowron A., Data Reduction Based on Rough Set Theory, *Proc. of The International Workshop on Statistics, Machine Learning and Knowledge Discovery in Databases*, Crete, Greece, April 28-29,1995
14. Skowron A., Polkowski L., Komorowski J. (1996). Learning Tolerance Relation by Boolean Descrip tions: Automatic feature extraction from data tabes. *Proc of The Fourth International Workshop on Rough Set, Fuzzy Set and Machine Discovery.* November 6-8,1996, Tokio, Japan, pp.11-17.
15. Skowron A., Stepaniuk J., Tolerance Approximation Spaces. *In Fundamenta Informaticae*, August 1996, Vol. 27, Numer 2,3, pp.245-253.
16. Windham, M.P. Geometric Fuzzy Clustering Algorithms, Fuzzy Set and Systems 3, pp.271-280, 1983.
17. Ziarko, W.P.(Ed.) Rough Set, Fuzzy Set and Knowledge Discovery, Springer-Verlag, London, 1994.

Appendix

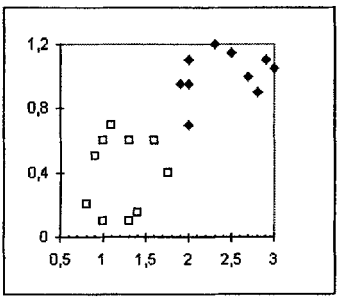

Fig. 1.a) The set of apples and pears

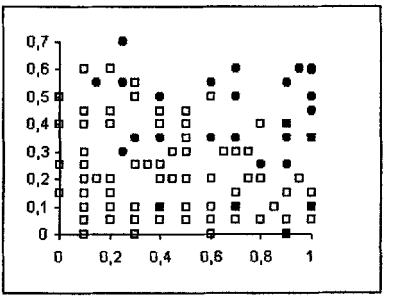

Fig. 1.b) The set of fruit pairs

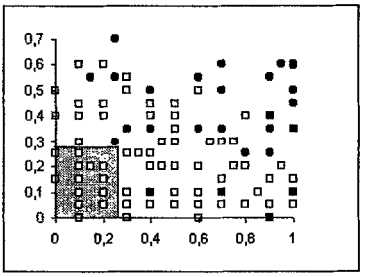

Fig. 2a) Tolerance relation $\mathcal{T}_1$

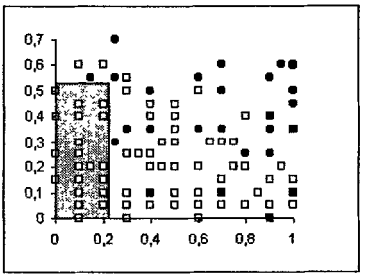

Fig. 2b) Tolerance relation $\mathcal{T}_2$

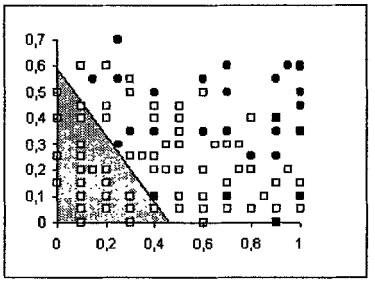

Fig. 3a) Tolerance relation $\mathcal{T}_3$

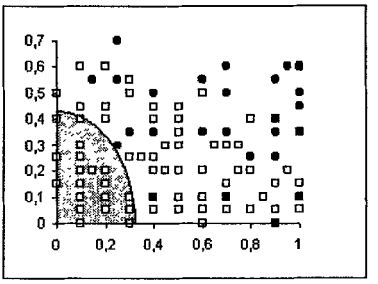

Fig. 3b) Tolerance relation $\mathcal{T}_4$

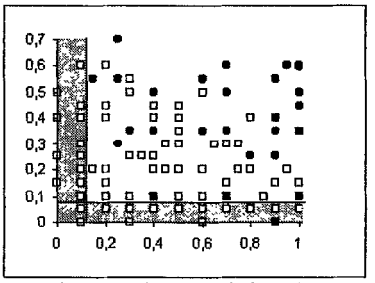

Fig. 4a) Tolerance relation $\mathcal{T}_6$

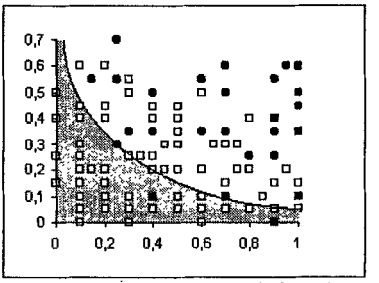

Fig. 4b) Tolerance relation $\mathcal{T}_7$