O P E R A T I O N S R E S E A R C H A N D D E C I S I O N S
No. 2 2017
DOI: 10.5277/ord170203

Leszek KLUKOWSKI[1]

DETERMINING AN ESTIMATE OF AN EQUIVALENCE RELATION FOR MODERATE AND LARGE SIZED SETS[2]

This paper presents two approaches to determining estimates of an equivalence relation on the basis of pairwise comparisons with random errors. Obtaining such an estimate requires the solution of a discrete programming problem which minimizes the sum of the differences between the form of the relation and the comparisons. The problem is NP hard and can be solved with the use of exact algorithms for sets of moderate size, i.e. about 50 elements. In the case of larger sets, i.e. at least 200 comparisons for each element, it is necessary to apply heuristic algorithms. The paper presents results (a statistical preprocessing), which enable us to determine the optimal or a near-optimal solution with acceptable computational cost. They include: the development of a statistical procedure producing comparisons with low probabilities of errors and a heuristic algorithm based on such comparisons. The proposed approach guarantees the applicability of such estimators for any size of set.

Keywords: *estimation of an equivalence relation, pairwise comparisons with random errors, concept of nearest adjoining order*

1. Introduction

The estimators of an equivalence relation based on multiple pairwise comparisons with random errors, proposed by Klukowski [9, 10], require the optimal solution of a discrete programming problem. Such a solution minimizes the difference between the form of a relation, determined in an appropriate way, and the comparisons. Under non-restrictive assumptions about errors in comparison, these estimates are consistent. The solution time is of exponential type [9] – as the number of pairwise comparisons in-

[1]Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warsaw, Poland, e-mail address: Leszek.Klukowski@ibspan.waw.pl

[2]The paper is an extended version of the text presented at BOS 2016 Conference, 13–14 October, Warsaw, Poland.

creases. Such optimization problems can be solved with the use of appropriate algorithms: complete enumeration – for sets including not more than several elements, discrete mathematical programming – up to 50 elements (assuming single comparisons of each pair), heuristic approaches – for sets exceeding 50 elements, especially in the case of multiple comparisons of each pair. Heuristic algorithms reduce computational costs, but can provide questionable solutions in the case when the probabilities of errors in comparisons are not close to zero. However, a large number of comparisons of any single element, i.e. at least 200 single comparisons or 100 multiple comparisons, can be advantageous. This is so, because such sized sets allow us to carry out some preprocessing and obtain new single comparisons with significantly reduced probabilities of errors. These comparisons can be generated with the use of the statistical procedure proposed in this paper. Such results can be used as the basis of the proposed heuristic algorithm and also as a starting point for an exact discrete algorithm. The computational cost of such a "combined" approach is typically acceptable. These features make the proposed approach, based on the concept of nearest adjoining order (Slater 1961), highly efficient and applicable for any size of set.

The paper consists of five sections. The second section presents the estimation problem, assumptions about pairwise comparisons and the form of the estimator. In the third section, we describe concisely some well-known optimization problems used to estimate an equivalence relation and suitable for sets with a moderate number of elements. The fourth section presents a statistical procedure generating pairwise comparisons with reduced probabilities of errors, based on a large number of initial comparisons, and the proposed algorithm. The last section summarizes the results.

2. Estimation problem, assumptions about comparisons, form of the estimator

2.1. Estimation problems

We are given a finite set of elements $\mathbf{X} = \{x_1, ..., x_m\}$ $(3 \le m < \infty)$. It is assumed that for the set $\mathbf{X}$ there exists an equivalence relation satisfying the conditions of reflexivity, transitivity and symmetry. This relation generates a family of subsets $\chi_1^*, ..., \chi_n^*$ $(n \ge 2)$, where each subset only includes equivalent elements.

The family $\chi_1^*, ..., \chi_n^*$ has the following properties:

$$\bigcup_{q=1}^{n} \chi_q^* = \mathbf{X} \tag{1}$$

$$\chi_r^* \cap \chi_s^* = \{\mathbf{0}\} \quad (r \neq s) \tag{2}$$

where: $\mathbf{0}$ – the empty set,

equivalent elements

$$x_i, x_j \in \chi_r^* \equiv x_i, x_j \tag{3}$$

non-equivalent elements

$$(x_i \in \chi_r^*) \wedge (x_j \in \chi_s^*) \equiv x_i, x_j \quad (i \neq j, r \neq s) \tag{4}$$

Any relation defined by (1)–(4) can be alternatively defined by the values $T(x_i, x_j)$ $((x_i, x_j) \in \mathbf{X} \times \mathbf{X})$, where

$$T(x_i, x_j) = \begin{cases} 0 \text{ if there exists } \chi_r^* \text{ such that } (x_i, x_j) \in \chi_r^* \\ 1 \text{ otherwise} \end{cases} \tag{5}$$

2.2. Assumptions about pairwise comparisons

The relation $\chi_1^*, ..., \chi_n^*$ is to be estimated on the basis of N ($N \geq 1$) comparisons of each pair $(x_i, x_j) \in \mathbf{X} \times \mathbf{X}$. Any comparison $g_k(x_i, x_j)$ $(k = 1, ..., N)$ estimates the actual value of $T(x_i, x_j)$, subject to disturbance by a random error.

The following assumptions are made:

A1. The number of subsets n is unknown.

A2. The probabilities of errors $g_k(x_i, x_j) - T(x_i, x_j)$ $(k = 1, ..., N)$ have to satisfy the following assumptions:

$$P(g_k(x_i, x_j) = T(x_i, x_j)) \geq 1 - \delta \quad (\delta \in (0, 1/2)) \tag{6}$$

$$P(g_k(x_i, x_j) = T(x_i, x_j)) + P(g_k(x_i, x_j) \neq T(x_i, x_j)) = 1 \tag{7}$$

A3. The comparisons $g_k(x_i, x_j)$ $((x_i, x_j) \in \mathbf{X} \times \mathbf{X};\ k = 1, ..., N)$ are independent random variables.

Assumptions A2–A3 reflect the following properties of distributions of errors in comparisons:

• the probability of correct comparison is greater than the probability of incorrect comparison (inequalities (6), (7)),

• zero is the median (in a "sharp" form) and mode of each distribution of the comparison error,

• the comparisons are realizations of independent random variables,

• the expected value of any error can differ from zero. Assumptions (1)–(3) are weaker than those commonly used in the literature (see [2]); they correspond, e.g., to the results of testing statistical hypotheses.

2.3. The form of the estimator

The estimator presented by Klukowski ([9], Chap. 3, [10]), is based on the sum of the absolute differences between the form of the relation (the values $T(x_i, x_j)$) and the comparisons $g_k(x_i, x_j)$ $((x_i, x_j) \in \mathbf{X} \times \mathbf{X})$. The estimates will be denoted by $\hat{\chi}_1, ..., \hat{\chi}_{\hat{n}}$ or $\hat{T}(x_i, x_j)$. They are obtained on the basis of the following discrete minimization problem:

$$\min_{\chi_r, ..., \chi_r \in F_{\mathbf{X}}} \left\{ \sum_{<i, j> \in R_m} \sum_{k=1}^{N} \left| g_k(x_i, x_j) - t(x_i, x_j) \right| \right\} \tag{8}$$

where: $F_{\mathbf{X}}$ – the feasible set: the family of all relations $\chi_1, ..., \chi_r$ in the set X, $t(x_i, x_j)$ – the values describing any relation $\{\chi_1, ..., \chi_r\}$ from $F_{\mathbf{X}}$, R_m – the set of the form $R_m = \{< i, j > \mid 1 \le i, j \le m; j > i\}$.

The estimate based on the objective function (8) may not be uniquely defined and the value of the function (8) is non-negative.

2.4. Properties of estimators

The analytical properties of such an estimator are based on the random variable: $\sum_{R_m} \sum_k \left| g_k(x_i, x_j) - T(x_i, x_j) \right|$. The following results have been obtained by Klukowski [9], where:

A. Expected values:

$$E\left(\sum_{R_m} \sum_k \left| g_k(x_i, x_j) - T(x_i, x_j) \right| \right) \text{ and } E\left(\sum_{R_m} \sum_k \left| g_k(x_i, x_j) - \tilde{T}(x_i, x_j) \right| \right)$$

i.e. the values corresponding to the actual and to any other relation $\tilde{T}(x_i, x_j)$, satisfy the inequality:

$$E\left(\sum_{<i,\ j>\in R_m}\sum_{k=1}^{N}\left|g_k(x_i,\ x_j)-T(x_i, x_j)\right|\right)<E\left(\sum_{<i,\ j>\in R_m}\sum_{k=1}^{N}\left|g_k(x_i,\ x_j)-\tilde{T}(x_i, x_j)\right|\right) \quad (9)$$

B. The variances of the above random variables divided by the number of comparisons N converge to zero, as $N\to\infty$, i.e.

$$\begin{aligned}\lim_{N\to\infty}\mathrm{Var}\left(\frac{1}{N}\sum_{<i,j>\in R_m}\sum_{k=1}^{N}\left|g_k(x_i, x_j)-T(x_i, x_j)\right|\right)=0\\ \lim_{N\to\infty}\mathrm{Var}\left(\frac{1}{N}\sum_{<i,j>\in R_m}\sum_{k=1}^{N}\left|g_k(x_i, x_j)-\tilde{T}(x_i, x_j)\right|\right)=0\end{aligned} \quad (10)$$

C. The probability of the inequality

$$\sum_{R_m}\sum_{k} g_k(x_i,\ x_j)-T(x_i,\ x_j)<\sum_{R_m}\sum_{k} g_k(x_i,\ x_j)-\tilde{T}(x_i,\ x_j)$$

being satisfied converges to one as $N\to\infty$, i.e.:

$$\lim_{N\to\infty}P\left(\sum_{<i,j>\in R_m}\left|\sum_{k=1}^{N}g_k(x_i, x_j)-T(x_i, x_j)\right|<\sum_{<i,j>\in R_m}\left|\sum_{k=1}^{N}g_k(x_i, x_j)-\tilde{T}(x_i, x_j)\right|\right)=1 \quad (11)$$

Moreover,

$$\begin{aligned}P\left(\sum_{<i,j>\in R_m}\sum_{k=1}^{N}\left|g_k(x_i, x_j)-T(x_i, x_j)\right|<\sum_{<i,j>\in R_m}\sum_{k=1}^{N}\left|g_k(x_i, x_j)-\tilde{T}(x_i, x_j)\right|\right)\\ \geq 1-\exp\left\{-2N\left(\frac{1}{2}-\delta\right)^2\right\}\end{aligned} \quad (12)$$

Inequality (12) is based on the Hoeffding ([6] inequality.

Relationships (A)–(C) guarantee consistency and fast convergence of the estimators to the actual relation.

3. Optimization problems defining estimates of the equivalence relation

Optimal solutions of the problem (8) can be obtained with the use of discrete optimization algorithms, also applied in the cluster analysis. They are usually formulated for a fixed number n (because methods exist for determining this number, see, e.g., [3], p. 3.5). Such discrete algorithms has been presented by numerous authors [5, 4, 1, 3, Chap. 3]).

An initial approach [11] has the form:

$$\min\left\{\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{l=1}^{m} d_{kl} z_{kj} z_{lj}\right\} \tag{13}$$

$$\sum_{j=1}^{n} z_{kj} = 1 \quad (k = 1, ..., m) \tag{14}$$

$$z_{kj} \in \{0, 1\} \quad (j = 1, ..., n, \ k = 1, ..., m) \tag{15}$$

where: d_{kl} – distance (dissimilarity) between elements x_k, x_l, z_{kj} – decision variable equals 1 if an element x_k is assigned to the j-th cluster, zero otherwise.

The problem (13)–(15) has a quadratic objective function, linear constraints and $\{0, 1\}$ variables. It can be applied for the case of single comparisons of each pair in the following way: the distances d_{kl} should be replaced by the comparisons $g_1(x_k, x_l)$ and the optimal solution z^*_{kj} determines the form of n subsets. The problem can also be applied in the case $N > 1$ by using the median from the comparisons $g_1(x_i, x_j), ..., g_N(x_i, x_j)$.

The problem (13)–(15) is hard to solve in its original form and, therefore, is linearized by assuming $y_{klj} = z_{kj} z_{lj}$ and adding the constraints $y_{klj} = z_{kj} + x_{lj} - 1$, $y_{klj} \le z_{kj}$, $y_{klj} \le z_{lj}$. This modified problem also has some drawbacks, especially the large number of variables. Therefore, other approaches have been proposed for the problem of estimating an equivalence relation [5, 4]:

$$\min\left\{\sum_{k=1}^{m-1}\sum_{l=k+1}^{m} d_{kl} z_{kl}\right\} \tag{16}$$

$$z_{kl} + z_{lq} - z_{kq} \le 1 \quad (k = 1, ..., m-2) \tag{17}$$

$$-z_{kl} + z_{lq} + z_{kq} \le 1 \ (l = k+1, \ ..., \ m-1) \tag{18}$$

$$z_{kl} - z_{lq} + z_{kq} \le 1 \quad (q = l+1, ..., m) \tag{19}$$

$$z_{kl} \in \{0, 1\} \quad (k = 1, ..., m-1; \ l = k+1, ..., m) \tag{20}$$

The optimization problem (16)–(20) can be solved with the use of dual linear relaxation and the revised simplex algorithm. However, this approach need not always provide an optimal solution and other approaches have also been developed [4, 5]. In general, they can be used when the number of elements is not (significantly) greater than 50.

4. Algorithm based on a procedure reducing the probabilities of errors

The problem (8) can be effectively solved with the use of heuristic algorithms in the case when the probabilities of errors in comparison are close to zero. Such probabilities indicate a low fraction of incorrect comparisons – the expected number of errors is equal to $(m(m-1)/2)\delta N$. A large number of elements, i.e., $m \ge 100$, together with multiple comparisons ($N > 1$) or $m \ge 200$, allows us to obtain "new" comparisons with probabilities of errors significantly lower than δ. The basis for such comparisons are statistical tests which infer the form of the distributions of parallel comparisons: $g_k(x_i, x_1)$ and $g_k(x_r, x_1), ..., g_k(x_i, x_m)$ and $g_k(x_r, x_m)$ $(k = 1, ..., N; r \ne i)$.

The null hypothesis has the form H_0: all the comparisons $g_k(x_i, x_j)$ and $g_k(x_r, x_j)$ $(k = 1, ..., N; \ r \ne i, j; \ i \ne j)$ have the same distributions, while under the alternative H_1: some of these comparisons have different distributions. These hypotheses can be replaced by: H_0: x_i, x_r are equivalent and H_1: x_i, x_r are not equivalent. The statistic proposed below is based on the values of the comparisons $g_k(x_i, x_j)$ and $g_k(x_r, x_j)$ $(k = 1, ..., N; \ r \ne i, j)$. For $(m-1)N \ge 200$, it has a Gaussian limiting distribution. This test allows us to estimate the probabilities of both types of errors. It is appropriate to assume that they take similar values. It is clear that such a test significantly reduces the probability of error, δ.

4.1. Test for the equivalency of elements

The proposed test is based on the random variables:

$$\eta_{irjk} = \begin{cases} 1 \text{ if } g_k(x_i, x_j) = g_k(x_r, x_j) \\ 0 \text{ if } g_k(x_i, x_j) \ne g_k(x_r, x_j) \end{cases} \quad (k = 1, ..., N; \ r \ne i, j) \tag{21}$$

The parameters of these (zero-one) variables are as follows: the expected value assumes, under H_0, the form:

$$E(\eta_{irjk}\,|H_0) = (1-\delta)^2 + \delta^2,\ (r \neq i, j;\ \ j \neq i) \tag{22}$$

the variance – the form:

$$\mathrm{Var}(\eta_{irjk}\,|H_0) = 2\delta(1 - 3\delta\ + 4\delta^2 - 2\delta^3) \tag{23}$$

If H_1 is true, and x_i equivalent to x_j or x_r equivalent to x_j, then the parameters of the variable η_{irjk} take the form:

$$E(\eta_{irjk}\,|\,H_1) = 2\delta(1-\delta) \text{ and } \mathrm{Var}(\eta_{irjk}\,|\,H_1) = 2\delta(1 - 3\delta\ + 4\delta^2 - 2\delta^3) \tag{24}$$

It is obvious that:

$$E(\eta_{irjk}\,|H_0) = (1-\delta)^2 + \delta^2 > E(\eta_{irjk}\,|\,H_1) = 2\delta(1-\delta) \tag{25}$$

and that the difference between these expressions is equal to: $1 - 4\delta(1-\delta)$.

The same parameters can be determined for the variables η_{irik} $(k = 1, ..., N)$, i.e., for $j = i$, assuming $g_k(x_i, x_i) \equiv 0$. They assume the form:

$$E(\eta_{irik}\,|H_0) = 1 - \delta \tag{26}$$

$$\mathrm{Var}\,(\eta_{irik}\,|H_0) = \delta(1-\delta) \tag{27}$$

$$E(\eta_{irik}\,|H_1) = \delta \tag{28}$$

$$\mathrm{Var}(\eta_{irik}\,|H_1) = \delta(1-\delta) \tag{29}$$

The variables η_{irik} have higher expected value and lower variance than the variables η_{irjk} $(j \neq i)$. The above results show that the expected values of the variables:

$$\frac{1}{(m-1)N}\sum_{r\neq i,\,j}\ \sum_{k=1}^{N} E(\eta_{irjk}\,|H_0) \text{ and } \frac{1}{(m-1)N}\sum_{r\neq i,\,j}\ \sum_{k=1}^{N} E(\eta_{irjk}\,|H_1) \tag{30}$$

are not the same: the expected value of the variable corresponding to H_1 is lower, while the variances of both variables are the same. Thus, the null hypothesis can be formulated in the form:

$$H_0 : \sum_{r\neq i,\, j} \sum_{k=1}^{N} E(\eta_{irjk}) = N(m-1)((1-\delta)^2 + \delta^2) + N(1-\delta) \tag{31}$$

while the alternative is given by:

$$H_1 : \sum_{r\neq i,\, j} \sum_{k=1}^{N} E(\eta_{irjk}) < N(m-1)((1-\delta^2) + \delta^2) + N(1-\delta) \tag{32}$$

The variance of both variables is equal to:

$$\begin{aligned} \mathrm{Var}\left(\sum_{r\neq i,\, j} \sum_{k=1}^{N} \eta_{irjk} \middle| H_0 \right) &= \mathrm{Var}\left(\sum_{r\neq i,\, j} \sum_{k=1}^{N} \eta_{irjk} \middle| H_1 \right) \\ &= 2(m-1)N\delta(1-3\delta+4\delta^2-2\delta^3) + N\delta\,(1-\delta) \end{aligned} \tag{33}$$

In the case of large mN, the hypotheses (31), (32) can be replaced by:

$$H_0' : \frac{1}{(m-1)N} \sum_{r\neq i,\, j} \sum_{k=1}^{N} E(\eta_{irjk}) = (1-\delta)^2 + \delta^2 \tag{34}$$

$$H_1' : \frac{1}{(m-1)N} \sum_{r\neq i,\, j} \sum_{k=1}^{N} E(\eta_{irjk}) < (1-\delta)^2 + \delta^2 \tag{35}$$

Under the null hypothesis, the test statistic has the following Gaussian distribution:

$$\begin{aligned} \text{exp. v.} &= (1-\delta)^2 + \delta^2 + \frac{1-\delta}{m-1} \\ \text{var} &= \frac{1}{(m-1)N} 2\delta \left(1-3\delta+4\delta^2-2\delta^3) + \frac{\delta(1-\delta)}{(m-1)^2 N} \right) \end{aligned} \tag{36}$$

The test has a one-sided rejection region, i.e., values lower than the critical value corresponding to the assumed significance level α.

An example

Let us examine the example: $\delta = 0.1$, $m = 100$, $N = 3$. The difference between the expected values of the individual statistics $E(\eta_{irjk} \middle| H_0) - E(\eta_{irjk} \middle| H_1)$ equals 0.64, the difference $E(\eta_{irik} \geq H_0) - E(\eta_{irik} \middle| H_1)$ equals 0.8, the variance of the distribution (36) is

equal to 0.0005 (standard deviation 0.02236). In the case of elements $x_i \in \chi_p^*$ and $x_r \in \chi_q^*$, $(i \neq r,\ p \neq q)$ included in different subsets, each with 10 elements, the difference between statistics (31) and (32) is equal to 0.1244. Therefore, a test based on the Gaussian distribution guarantees that both probabilities of error are lower than 0.003 and the expected number of incorrect comparisons is lower than 15 (the total number of comparisons is equal to 4550).

The example shows that any significant reduction in the probability of an error δ requires the subsets $\chi_1^*, ..., \chi_n^*$ to be of the appropriate size. Typically, the minimal size should be at least several percent of the number m. However, before estimation, these sizes are not known. Therefore, it is suggested that some minimal size $\chi_{\min}^*$ is assumed to guarantee the necessary reduction in the probability δ (this can be done on the basis of the distribution of the test statistic (36)). Next we should detect and exclude elements of subsets of size lower than $\chi_{\min}^*$. These elements may be associated with an estimate based on a reduced set of equivalence groups, as a next step.

The detection of "small" subsets can also be done based on a statistical test. The null hypothesis assumes the following form: $\sum_{j \neq i} T(x_i,\ x_j) = m - \nu - 1$ $(i = 1, ..., m)$. Under the alternative, $\sum_{j \neq i} T(x_i, x_j) < m - \nu - 1$, where: ν is a natural number guaranteeing the required reduction in the probability of error δ; typically $\nu \leq \zeta(m-1)$, where $\zeta = 0.05$. This test can be based on the statistic: $(1/N)\sum_{j \neq i}\sum_{k=1}^{N} g_k(x_i,\ x_j)$. Its expected value and variance can be determined under the null hypothesis; they are equal to $(m-1-\nu)(1-\delta)+\nu\delta$ and $(m-1)\delta(1-\delta)/N$, respectively. Under the alternative, the expected value is lower than $(m-1-\nu)(1-\delta)+\nu\delta$, the variance is the same. In the case $mN \geq 200$, the Gaussian asymptotic distribution can be applied. Rejecting the null hypothesis for an element x_i means that it does not belong to a small subset. Rejecting it for the whole set **X** indicates a lack of small subsets. The opposite result, acceptance of H_0 – indicates inclusion into a small subset.

The comparisons obtained after the above preprocessing (with low probabilities of errors and without small subsets) are satisfactory for heuristic algorithms performing the partitioning or agglomeration of elements. The algorithm proposed below belongs to the second group.

4.2. The form of the algorithm

The comparisons obtained on the basis of the hypotheses H_0 and H_1 are denoted $\Gamma = \gamma(x_i, x_j)\ (< i,\ j > \in R_m)$. The result $\gamma(x_i, x_j) = 0$ corresponds to H_0, while

$\gamma(x_i,\ x_j)=1$ – to H_1. The comparisons $\gamma(x_i,x_j)$ allow us to infer for each element $x_i \in \mathbf{X}$, two sets: the former one, $\Psi(x_i)$, comprises the indexes of equivalent elements (corresponding to acceptance of H_0), the latter, $\Omega(x_i)$, the indexes of non-equivalent elements (corresponding to acceptance of H_1). It is clear that for equivalent elements x_i, x_j we have $\Omega(x_i)=\Omega(x_j)$. The sets $\Psi(x_i),\Psi(x_j)$ satisfy the relationship $\Psi(x_i)-\{j\}=\Psi(x_j)-\{i\}$. Thus, the algorithm minimizing the function (8) can be based on detecting subsets $\hat{\chi}_r$ $(r=1,...,\hat{n})$ with these features or similar features.

START

1. Exclude "small" subsets from the estimated relation and determine the probabilities of errors for the test with hypotheses (31), (32).

Test the null hypothesis H_0 for $(x_i, x_j) \in \mathbf{X}\times\mathbf{X}$ against the alternative H_1 (32) on the basis of comparisons $g_k(x_i, x_j)$, $g_k(x_r, x_j)$ $(k=1,\ ...,\ N;\ r\neq i,\ j)$ or assumed probabilities of errors (the results $\Gamma=\gamma(x_i, x_j)$ $(i=1,\ ...,\ m,\ j\neq i)$). Determine the upper limit m_d of the difference $\sum_{j\neq i}\left|\gamma(x_i, x_j)-\gamma(x_r, x_j)\right|$:

$$m_d=\operatorname{int}\left[2\alpha(1-\alpha)(m-1)+3\left((2\alpha(1-\alpha)(1-2\alpha(1-\alpha))(m-1)\right)\right)^{0,5}+0.5\right]$$

where: α – significance level when testing H_0, int[z] – integer part of z.

2. Create (non-overlapping) subsets $\breve{\chi}_s$ $(s=1,\ ...,\ \breve{n})$ from the elements of the set $\mathbf{X}$ with the following property: $\sum_{j\neq i}\left|\gamma(x_i,\ x_j)-\gamma(x_r,\ x_j)\right|\leq m_d$, for each $x_i,\ x_r \in \breve{\chi}_s$. Determine the value of the objective function (8) after this operation. This value is denoted F_{cur}.

3. Determine the set Δ, comprising elements of the set $\mathbf{X}$ with a significant contribution to the value of the objective function F_{cur}, i.e. all the elements x_i satisfying the inequality:

$$\sum_{j\neq i}\left|\breve{t}(x_i,\ x_j)-\gamma(x_i,\ x_j)\right|>m_h$$

where:

$$m_h=(m-1)\alpha+3\left(\alpha(1-\alpha)(m-1)\right)^{0.5}$$

If the set Δ is empty $(\#\Delta=0)$ go to 5.

4. Determine the best relocation for each element of the set Δ, i.e. into a subset $\breve{\chi}_q$ $(1 \le q \le \breve{n})$ or a new subset $\breve{\chi}_{\breve{n}+1}$ which achieves the maximal decrease in the function (8). Perform these relocations starting from an element corresponding to the maximal decrease in the function (8). If the value F_{cur} has decreased in this step, return to 3.

5. Accept $\breve{\chi}_q$ $(1 \le q \le \breve{n})$ as the estimate $\hat{\chi}_q$ $(q = 1, ..., \hat{n})$.

END

The above algorithm is composed of two phases. The former phase involves agglomerating of all elements x_i, x_r with similar sets $\Omega(\cdot)$, $\Psi(\cdot)$, i.e. satisfying the inequality: $\sum_{j \ne i} \left| \gamma(x_i, x_j) - \gamma(x_r, x_j) \right| \le m_d$. The value m_d is determined as the sum comprising: the expected value of the variable $\sum_{j \ne i} \left| \gamma(x_i, x_j) - \gamma(x_r, x_j) \right|$ $((x_i, x_r) \in \chi_q^*$ $(1 \le q \le n))$ and its three standard deviations, assuming binomial distribution of the sum (in fact some of its components can be not independent random variables, but three standard deviations "compensate" this fact). This phase ends when each subset $\breve{\chi}_s$ $(1 \le s \le \breve{n})$ includes elements satisfying the inequality $\sum_{j \ne i} \left| \gamma(x_i, x_j) - \gamma(x_r, x_j) \right| \le m_d$. Such a partition can be not unique; therefore objective function indicates the best result.

The second phase is oriented at "improving" the estimate obtained in the first phase. The elements $x_i \in \mathbf{X}$ of the current estimate $\breve{\chi}_s$ $(1 \le s \le \breve{n})$ which have a significant contribution to the objective function (8) are detected. The threshold value of the contribution m_h is determined on the basis of expected value of the random variable $\left(\sum_{j \ne i} \left| \breve{t}(x_i, x_j) - \gamma(x_i, x_j) \right| \right)$ $((x_i, x_r) \in \breve{\chi}_q$ $(1 \le q \le \breve{n}))$ and its three standard deviations determined under assumption that $\breve{t}(x_i, x_j) = T(x_i, x_j)$. The elements with a significant contribution $\sum_{j \ne i} \left| \breve{t}(x_i, x_j) - \gamma(x_i, x_j) \right|$ are relocated to subsets, which leads to a decrease in the value of the objective function (8). This phase ends after no such elements remain.

An estimate corresponding to the value of the objective function being equal to zero gives an exact optimal solution, while those with low values – can be close to exact or even exact. It is clear that comparisons $\gamma(x_i, x_j)$ $((x_i, x_j) \in \mathbf{X} \times \mathbf{X})$ which have very low probabilities of errors (not greater than 10^{-3}) are also useful for discrete programming algorithms for sets **X** with more than 50 elements. The computational cost may be acceptable in this case.

The literature on this subject contains many other heuristic algorithms (see, e.g., [5]). The estimate obtained in such a way can be verified by testing the hypothesis stating that

such a relation exists against the hypothesis that comparisons are completely random or all the elements are equivalent (see, e.g., [9, 3, Chapt. 7]). Verifying the existence of individual subsets $\hat{\chi}_r$ $(1 \le r \le \hat{n})$ can be done with the use of, e.g., the Cochran test.

5. Concluding remarks

An algorithm for solving the optimization problem has been presented aimed at obtaining estimates of an equivalence relation on the basis of pairwise comparisons with random errors. The objective function of this problem expresses the difference between the form of the relation and the comparisons. Such an approach is applicable for moderately sized (about 50 elements) and large sets (at least 100 elements with multiple comparisons). Cases with a moderately sized set can be solved with the use of well-known exact algorithms. When a large number of comparisons are made, another approach is recommended, which allows the construction of tests generating "new" comparisons with significantly reduced probabilities of errors. Such comparisons enable application of the heuristic algorithm proposed in this paper. The results obtained from such an algorithm can give a final estimate, if the value of the criterion function is equal or close to zero, or provides a starting point for exact algorithms. This algorithm performs nearly perfectly when the probabilities of errors in comparisons are low (below 0.01) and subsets are appropriately sized χ_q^* $(1 \le q \le n)$. This results from the fact that the value $E\left(\sum_{j \ne i} \left|\gamma(x_r, x_j) - \gamma(x_r, x_j)\right|\right)$ is significantly different in the cases $(x_r, x_j) \in \chi_q^*$ and $(x_r, x_j) \notin \chi_q^*$. Thus, an approach based on minimizing the differences between the comparisons and the form of the relation is useful, computationally efficient and reliable for any size of set. It should be emphasized that the statistical properties of estimates based on minimizing the function (8) have been determined [9] and can be verified by the use of statistical tests. Moreover, their precision is also evaluated by the value of the objective function (8).

References

[1] CHOPRA R., RAO M.R., *The partition problem*, Math. Progr., 1993, 59, 87–115.
[2] DAVID H.A., *The Method of Paired Comparisons*, 2nd Ed., Griffin, London 1988.
[3] GORDON A.D., *Classification*, 2nd Ed., Chapman and Hall CRC, 1999.
[4] HANSEN P., JAUMARD B., *Cluster analysis and mathematical programming*, Math. Progr., 1997, 79, 191–215.

[5] HANSEN P., JAUMARD B., SANLAVILLE E., *Partitioning Problems in Cluster Analysis. A Review of Mathematical Programming Approaches. Studies in Classification, Data Analysis and Knowledge Organization*, Springer-Verlag, 1994.
[6] HOEFFDING W., *Probability inequalities for sums of bounded random variables*, J. Am. Stat. As., 1963, 58, 13–30.
[7] KLUKOWSKI L., *Some probabilistic properties of the nearest adjoining order method and its extensions*, Ann. Oper. Res., 1994, 51, 241–261.
[8] KLUKOWSKI L., *The nearest adjoining order method for pairwise comparisons in the form of difference of ranks*, Ann. Oper. Res., 2000, 97, 357–378.
[9] KLUKOWSKI L., *Methods of estimation of relations of: equivalence, tolerance, and preference in a finite set*, IBS PAN, Ser. Systems Research, Vol. 69, Warsaw 2011.
[10] KLUKOWSKI L., *Estimators of the relations of equivalence, tolerance and preference based on pairwise comparisons with random errors*, Oper. Res. Dec., 2012, 22, 15–34.
[11] RAO M.R., *Cluster analysis and mathematical programming*, J. Am. Stat. As., 1971, 66, 622–626.
[12] SLATER P., *Inconsistencies in a schedule of paired comparisons*, Biometrika, 1961, 48, 303–312.

Received 4 November 2016
Accepted 13 April 2017