Panel matrix and ranking model recovery using mixed-scale measured data

O. Y. Bakhteev
bakhteev@phystech.edu
Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Russia

A decision-making problem is solved in the field of operational research education. The paper presents a method for recovery of changes in ratings of student employees. These ratings are based on interviews at the information technology (IT) training center. A dataset consisting of expert estimates for assessments for different years and overall rating for these students is considered. The scales of the expert estimates vary from year to year, but the scale of the rating remains stable. One should recover the time-independent ranking model. The problem is stated as the object–feature–year panel matrix recovery. It is a map from student descriptions (or their generalized portraits) to expected ratings for all years. Also, a stability of the ranking model produced by the panel matrix is studied. A new method of panel matrix recovery is suggested. It is based on a solution of multidimensional assignment problem. To construct a ranking model, an ordinal classification algorithm with partially ordered feature sets and an algorithm based on support vector machine have been used. The problem is illustrated by the dataset containing the expert assessment of the student interviews at the IT center.

Keywords: *operational research education; business analytics; knowledge extraction; ratings; expert estimates; clustering; mixed scales*

DOI: 10.21469/22233792.1.14.05

1 Introduction

The paper presents a solution for the panel matrix recovery problem, where the panel matrix is a multidimensional object–feature–year [1] matrix. The objects of the matrix are represented by vectors containing different object features for several years. This algebraic structure is used to recover the ranking model and estimate its stability: whenever the parameters of the model remain stable in different years, is considered to be stable. The original dataset is represented by the design matrix, namely, the object–feature matrix, which contains all the object descriptions during all the timestamps.

The main goal of this paper is to develop an algorithm of panel matrix recovery and to recover the ranking model. Let the panel matrix $\mathbf{Z}$ be the matrix, where the entry z_{ijt} is the feature j of the student i in the year t.

The problem of the panel matrix recovery can be found in the pattern recognition, when it is required to recover the tracks of different targets received by sensors [2]. In this paper, another application problem is considered that can be met in business-analytics: an employee selection problem. The dataset containing expert assessments, which were received during the interview at an educational IT-center in 2006–2009, is considered. The purpose is to recover the ranking model and to estimate the stability of this ranking model during all the years. It is proposed to construct some generalized "portraits" of these students and to recover the panel matrix $\mathbf{Z}$ based on these portraits. Note that in this paper, a special case of the panel matrix recovery is considered when the features (answers from assessment) of the portraits remain stable and the only elements that are changeable are the classes of students. The scheme of the panel matrix recovery is shown in Fig. 1.

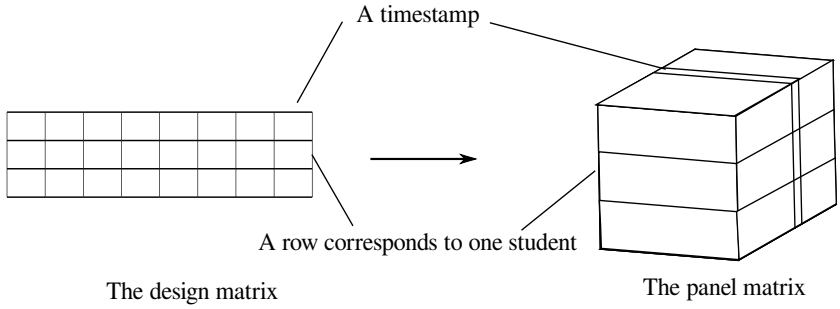

Figure 1 The panel matrix recovery. The generalized student "portraits" that remain stable during all the time are found and considered to be the panel matrix objects

The problem is stated as the multidimensional assignment problem. It requires to find a bijection between object descriptions in different years. The main difficulty in solving this problem is that the multidimensional assignment problem is NP-hard (nondeterministic polynomial-time hard) [3]; therefore, it requires to use heuristic algorithms to solve it. There are several solutions for this problem and related problems [4–6]. The papers [3, 7] propose to use linear programming and randomization algorithms. The methods proposed in the present paper are based on a hypergraph construction. One can use a genetic algorithm [8]. As an alternative, the problem is stated as the common min-cost max flow problem [9].

Define some terms that will be used for the dataset description.

Definition 1. A scale $\mathbb{L}$ is an algebraic structure [10] with a fixed set of operations, relations, and a fixed set of axioms.

Definition 2. A nominal scale $\mathbb{C}$ is a scale with a fixed binary relation:

1) $x = y \vee x \neq y$;
2) $x, y : x = y \Rightarrow y = x$; and
3) $x, y, z : x = y \wedge y = z \Rightarrow x = z$

where x, y, and z are the objects from the scale $\mathbb{C}$: $x, y, z \in \mathbb{C}$.

Definition 3. An oridnal scale $\mathbb{O}$ is a nominal scale with a fixed relation:

1) xRx;
2) $xRy \wedge yRx \Rightarrow x = y$; and
3) $xRy \wedge yRz \Rightarrow xRz$

where $x, y, z \in \mathbb{O}$.

Definition 4. A linear scale $\mathbb{W}$ is an ordinal scale with total order and addition and subtraction operations defined on it.

Definition 5. A ranking function f is a mapping from the object space $\mathbb{X}$ to the finite set of classes $\mathbb{Y}$ with a total order defined on it [11].

The ranking model recovery problem can be met not only in the employee selection but also in information technologies [12], agriculture [13], and energy management [14]. The type of the ranking model recovery algorithm can be chosen with respect to the dataset scale [15–17]. In this paper, a pairwise dominating matrix algorithm is considered for the feature set with a partial order defined on each feature [18]. Another algorithm considered in the present paper

is an algorithm RankSVM, which is a generalization of a classification algorithm based on the support vector machine (SVM) [19].

The ranking model is recovered by the dataset [20] containing students that attempt to pass the interview at the educational center during 2006–2009. The data can contain missing values. The dataset feature descriptions are shown in Table 1.

The expert proposes that each feature should give a positive contribution into the rating. The higher score student gets the higher rating he receives according to the "bigger is better" [21] principle. The nominal feature "Student's interests" is not used in the ranking model recovery, but it is used in the panel matrix recovery in order to cluster students. The expert also recommends to round the feature "Student's interests" in order to get three discrete values.

One of the steps of the panel matrix recovery is a clustering, which requires a distance function. This function determines how close to each other the students estimates are. A generalized Heterogeous Euclidean-Overlap Metric (HEOM) function [22] and Heterogeous Manhattan-Overlap Metric (HMOM) function [23] are proposed for a mixed-scale dataset (a dataset containing linear, ordinal [24], and nominal scales). Extracting significant information from such datasets is a challenging high priority issue for many organizations in the business analytics.

2 The problem formulation

In this section, a formal definition of the panel matrix $\mathbf{Z}$ and ranking model recovery problem are presented.

Definition 6. The panel matrix $\mathbf{Z}$ is a matrix where the entry z_{ijt} is the feature j (answers from assessment) of student i in year t.

The dataset contains the set of pairs of mixed-scale data:

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i) : i \in \mathcal{I}\}, \quad \text{the object index } i \in \mathcal{I} = \{1, \ldots, m\},$$

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^{\mathrm{T}}, \quad y_i \in \mathbf{y}$$

with metric

$$d : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+$$

Table 1 Dataset feature descriptions

Feature	Scale type	Scale cardinality
Average score during university education	Linear, $\mathbb{W}$	Rational number in [3;5]
Average score for the last term		
Acceptance preference (expert estimation)	Ordinal, $\mathbb{O}$	Rational number, the cardinality changes during some years
Student's interests: programming, telecommunication development, or both	Nominal, $\mathbb{C}$	The experts used 3 discrete values {programming, both, telecommunication} in 2006; later, the experts used rational number
Students' responsibility	Ordinal, $\mathbb{O}$	Rational number, the cardinality changes during some years
Level of knowledge		
Motivation		
Student's class — the final rating in the assessment		

where $\mathbb{X} = \mathbb{L}_1 \times \cdots \times \mathbb{L}_n$ is the object space; $\mathbf{X}$ is the object–feature matrix for the dataset; $\mathbf{x}_i \subset \mathbb{X}$; and $\mathbf{y}$ is the vector of classes for each object in dataset such that its elements are in $\mathbb{Y}$. In this paper, the generalized HEOM distance and HMOM distance functions are used as the functions of d (see Eqs. (12) and (14) below). Define a total order on the set of classes:

$$\mathbb{Y} = \{\text{“1”}, \text{“2”}, \text{“3”}, \text{“4”}, \text{“5”}\} \tag{1}$$

where “1” $\prec$ “2” $\prec$ “3” $\prec$ “4” $\prec$ “5”.

Let $\mathcal{T} = \{t\}$ be the set of timestamps of the estimations. In this paper, the set $\mathcal{T}$ contains 4 elements, corresponding to 2006–2009. Let $\mathbf{X}^t$ be the matrix of the objects $\mathbf{X}$ of the year t. Let $\mathbf{D}^t$ be the distance matrix for all pairs of objects per year t:

$$d_{iq}^t = d(\mathbf{x}_i^t, \mathbf{x}_q^t), \quad \mathbf{x}_i^t, \mathbf{x}_q^t \in \mathbf{X}^t.$$

The panel matrix recovery procedure consists of the following parts:

11) a dendrogram constructing algorithm $\mathfrak{d}$;
21) a clustering algorithm $\mathfrak{c}$;
31) a class recovery algorithm $\mathfrak{r}$;
41) a bijection recovery algorithm $\mathfrak{m}$ that finds a bijection between cluster centroids $\mathbf{M}^t$ of different years; and
51) an algorithm $\mathfrak{a}$ of averaging cluster centroids.

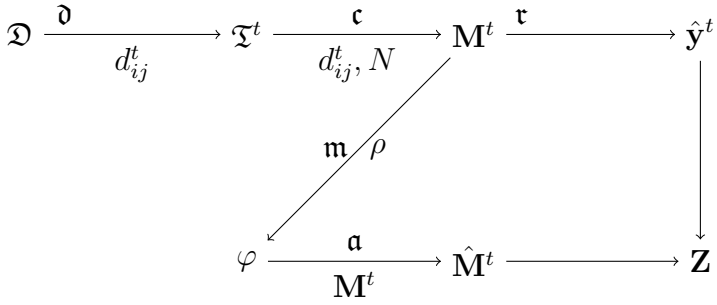

Figure 2 Panel matrix recovery procedure

The panel matrix recovery procedure is shown in Fig. 2: for each year t, the algorithm $\mathfrak{d}$ constructs the dendrogram $\mathfrak{T}^t$. Then, calculate the optimal number of clusters N and the algorithm $\mathfrak{c}$ proceeds clustering. For each cluster $\boldsymbol{\mu}$ from the set of cluster centroids $\mathbf{M}^t$, the algorithm $\mathfrak{r}$ recovers its class $\hat{y}^t \in \hat{\mathbf{y}}^t$. After that, the algorithm $\mathfrak{m}$ finds a bijection φ that matches clusters from different years $\mathbb{Y}$. As a result, get the panel matrix $\mathbf{Z}$ from the averaged centroids $\hat{\mathbf{M}}$, which correspond to the student portraits, and the vector of recovered classes $\hat{\mathbf{y}}^t$.

The algorithm $\mathfrak{c}$ clusters the objects of the dataset for each year t. Let $\mathbf{M}^t \subset \mathbb{X}$ be the set of N cluster centroids for year t, $\mathbf{M}^t = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N]^T$.

For each cluster centroid $\boldsymbol{\mu}_k^t$, recover its class $\hat{y_k}^t \in \mathbb{Y}$ (1). Here, for this purpose, median function has been used:

$$\hat{y}_k^t = \text{median}\{y_i^t : \text{cluster}(\mathbf{x}_i^t) = k\}$$

where $\text{cluster}(\mathbf{x})$ is the function which returns the index of cluster that contains element $\mathbf{x}$.

Let the distance function be given by

$$\rho : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+. \tag{2}$$

This function is used in algorithm $\mathfrak{m}$ to find the mapping that satisfies criteria (5) and (4) (see below). The distance function used as the function of ρ is also described below (15).

The algorithm $\mathfrak{m}$ of the bijection recovery between clusters of different years finds the permutation of cluster indexes:

$$\varphi : \{1, \ldots, N\} \to \{1, \ldots, N\} \tag{3}$$

such that for each year t the mapping is a bijection. Let use the distance function ρ (2) to find this mapping. A set of cluster centroids is called $\mathbf{G}_k = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{|\mathcal{T}|}]$ if it contains all the centroids that φ returns k for them:

$$\mathbf{G}_k = \{\boldsymbol{\mu} \in \mathbb{X} : \varphi(\text{index}(\boldsymbol{\mu})) = k\}$$

where $\text{index} : \mathbf{M}^t \to \{1, \ldots, N\}$ is the function, which returns the index for each cluster.

Let us select φ that minimizes the following criteria:

1. The clustering criterion $C_{\mathcal{C}}$: φ should minimize the average value of R where R is the ratio from the average distance between objects $\boldsymbol{\mu}_{k_1}$ and $\boldsymbol{\mu}_{k_2}$ from cluster set $\mathbf{G}_k$ to the average distance between cluster centroids $\mathbf{G}_{k_1}$ and $\mathbf{G}_{k_2}$:

$$C_{\mathcal{C}} = \text{mean}_{k \in \{1,\ldots,N\}} R(\mathbf{G}_k), \quad R(\mathbf{G}_k) = \frac{\text{mean}_{\boldsymbol{\mu}_{k_1}, \boldsymbol{\mu}_{k_2} \in \mathbf{G}_k} d(\boldsymbol{\mu}_{k_1}, \boldsymbol{\mu}_{k_2})}{\text{mean}_{\mathbf{G}_{k_1}, \mathbf{G}_{k_2}} d(\mathbf{G}_{k_1}, \mathbf{G}_{k_2})}. \tag{4}$$

2. The stability criterion $C_{\mathcal{S}}$: φ should minimize the difference in classes $\hat{y}_{k_1}$ and $\hat{y}_{k_2}$ of cluster set $\mathbf{G}_k$:

$$C_{\mathcal{S}} = \sum_{k=1}^{N} \sum_{\boldsymbol{\mu}_{k_1}, \boldsymbol{\mu}_{k_2} \in \mathbf{G}_k} |\hat{y}_{k_1} - \hat{y}_{k_2}|. \tag{5}$$

The resulting optimization problem is the following:

$$\begin{cases} \varphi = \arg\min\limits_{\varphi' \in \Phi} C_{\mathcal{C}}; \\ \varphi = \arg\min\limits_{\varphi' \in \Phi} C_{\mathcal{S}} \end{cases}$$

where Φ is the set of mappings from the index set $\{1, \ldots, N\}$ to itself such that for each year t, the mapping is bijective.

As an averaging algorithm $\mathfrak{a}$, one gets averaged cluster centroids using the following function:

$$\text{avg}\{\hat{\boldsymbol{\mu}}_{kj}\} = \begin{cases} \text{mean}\{\mu_{qj} : \boldsymbol{\mu}_q \in \boldsymbol{G}_k\} & \text{whenever } \mathbb{L}_j \text{ is linear scale;} \\ \text{median}\{\mu_{qj} : \boldsymbol{\mu}_q \in \boldsymbol{G}_k\} & \text{whenever } \mathbb{L}_j \text{ is ordered scale;} \\ \text{mode}\{\mu_{qj} : \boldsymbol{\mu}_q \in \boldsymbol{G}_k\} & \text{whenever } \mathbb{L}_j \text{ is nominal scale} \end{cases} \tag{6}$$

where $\hat{\boldsymbol{\mu}} \in \hat{\mathbf{M}}$ is the averaged cluster centroid from G_k; and $\hat{\mathbf{M}}$ is the set of averaged cluster centoids. Let use $\hat{\mathbf{M}}$ as an object set for the panel matrix $\mathbf{Z}$.

As a result of the panel matrix recovery procedure, obtain the matrix $\mathbf{Z}$ that contains the set of the averaged centroids $\hat{\mathbf{M}}$ and the vector of recovered classes $\hat{y}_i^t \in \mathbf{y}^t$.

Ranking model recovery

To solve the ranking model recovery problem, one should find a mapping:

$$f : \mathbb{X} \to \mathbb{Y}$$

which minimizes error function $Q(\mathbf{X})$. In this paper, Kendall correlation coefficient [25] has been used:

$$Q(\mathbf{X}) = 1 - \text{KendallTau}(\mathbf{y}, \hat{\mathbf{y}})$$

where $\hat{\mathbf{y}}$ is the vector of classes which is returned for objects $\mathbf{X}$ by the function f; and the Kendall correlation coefficient is

$$\text{KendallTau} = \frac{4|\{(i,q) : y_i > y_q, \hat{y}_i > \hat{y}_q\}|}{m(m-1)} - 1. \tag{7}$$

3 Calculating optimal number of clusters

In the previous section, the number of clusters N was considered to be fixed. One can select the value for N using expert estimates. The other way is to optimize the number of clusters using heuristics. This section describes the optimization problem, which can be used as the one way to find the optimal number of clusters N. Assume the number N of clusters remains stable for each year from the set $\mathcal{T}$. The reason of this assumption is the wish to recover the ranking model for each year of the panel matrix and to estimate correlation between rankings of different years. If the number N differs for different years, this problem is incorrect.

Optimize the number of clusters N using dendrogram constructing algorithm $\mathfrak{d}$.

Definition 7. The dendrogram $\mathfrak{T}^t$ is a tree that is built using the distance matrix $\mathbf{D}^t$ which shows the relationships between clusters.

Describe the dendrogram constructing method. Suppose one has a linkage algorithm:

$$A_{\mathcal{L}} : {\mathbb{R}_+}^m \times {\mathbb{R}_+}^n \to \mathbb{X} \times \mathbb{X}\,. \tag{8}$$

It defines the pair of elements $\mathbf{x}_i$ and $\mathbf{x}_q$ to merge into one cluster $\boldsymbol{\mu}_k$. Let merge this pair and then recalculate the distance matrix $\mathbf{D}^t$ using information about the merged elements.

At the end of dendrogram constructing algorithm, one receives a tree $\mathfrak{T}^t$. Its root contains two last elements merged at the final step.

The example of dendrogram is shown in Fig. 3. The elements A, B, and C are clustering until one cluster remains.

Theorem 1. For each $N \in \{1, \ldots, m\}$, a clustering with a set of N clusters is constructible, where m is the number of objects.

Proof. Each step, the number of clusters is reduced by one. Therefore, after $m - N$ steps, one gets the set of clusters with cardinality equal to N. ∎

Let us construct the dendrogram $\mathfrak{T}^t$ for each year t for optimal N calculating. The number of clusters N is optimal whenever it satisfies the following criteria.

1. The uniform class criterion $C_{\mathcal{U}}$: the number of cluster centroids $\mathbf{M}^t$ of different classes should be equal. N should minimize the deviation of number of different classes of clusters:

$$C_{\mathcal{U}}(\mathbf{M}^t) = \sigma\{|\mathbf{M}^t_y|, y \in \mathbb{Y}\}$$

where $|\mathbf{M}^t_i|$ is the cardinality of the set $\mathbf{M}^t$ with class $y \in \mathbb{Y}$; and σ is the standard deviation.

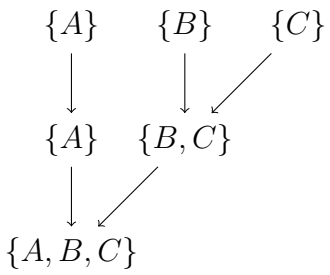

Figure 3 The example of dendrogram

2. Mixing class criterion $C_{\mathcal{M}}$: the number of clusters N should decrease the difference of classes inside clusters:

$$C_{\mathcal{M}}(\mathbf{M}^t) = \text{mean}_{\boldsymbol{\mu}_k \in \mathbf{M}^t} \sigma(\{y_i : \text{cluster}(\mathbf{x}_i) = k\}).$$

The number of clusters N should be less than or equal to the minimum number of objects in the sets: $N \leqslant \min\limits_{t \in \mathcal{T}} |\mathbf{X}^t|$. Also, let construct a clustering that contains a representative of each class; therefore, N should be greater than or equal to the cardinality of $\mathbb{Y}$. The final formula for the optimization problem is the following:

$$\begin{cases} N = \arg\min\limits_{N}(\text{mean}_{t \in \mathcal{T}}(\delta_{\mathcal{U}}(\mathbf{M}^t))); \\ N = \arg\min\limits_{N}(\text{mean}_{t \in \mathcal{T}}(\delta_{\mathcal{M}}(\mathbf{M}^t))); \\ N \geqslant 5, \quad N \leqslant \min\limits_{t \in \mathcal{T}} |\mathbf{X}^t|. \end{cases} \tag{9}$$

Some heuristics have been proposed to select N. Let us construct two dendrograms $\mathfrak{T}^t_{\mathcal{U}}$ for each year t. They use the linkage algorithms (8) $A_{\mathcal{LE}}$ and $A_{\mathcal{LM}}$ to estimate functionals δ_Y and δ_E.

In order to estimate $C_{\mathcal{U}}$, let us use the following linkage algorithm:

$$A_{\mathcal{LU}} = \underset{\substack{\boldsymbol{\mu}_{k_1}, \boldsymbol{\mu}_{k_2} \in \mathbf{M}^t, \\ \hat{y}_{k_1} = \hat{y}_{k_2} = \max\limits_{i = \{1,\ldots,5\}} |\mathbf{M}^t_i|}}{\arg\min} D_{k_1 k_2}.$$

Select a pair of the closest objects of the most common class (the class which has the most number of representatives). Each step, the cardinality of the largest set $\mathbf{M}^t_i$ of cluster centroids of the fixed class y has been reduced. The difference in cardinality between these sets decreases. Therefore, the dendrogram $\mathfrak{T}^t_E$ is quite close to be optimal with respect to δ_Y for the fixed N.

In order to estimate $C_{\mathcal{M}}$, the following linkage algorithm has been used:

$$A_{\mathcal{LM}} = \underset{\substack{\boldsymbol{\mu}_{k_1}, \boldsymbol{\mu}_{k_2} \in \mathbf{M}^t, \\ |\hat{y}_{k_1} - \hat{y}_{k_2}| = \max}}{\arg\min} ||\text{members}(\boldsymbol{\mu}_{k_1})| - |\text{members}(\boldsymbol{\mu}_{k_2})||_2,$$

where members $\mathbf{M}^t \to 2^{\mathbf{X}^t}$ are the functions that return a set of objects assigned to the cluster. This linkage algorithm selects the pair $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ of clusters with the largest difference in classes and with the smallest difference in cardinality. Each step, the difference in classes inside some

cluster has been maximized; therefore, the dendrogram is quite close to be the worst with respect to δ_M for the fixed N.

To find a compromise between two criteria $C_{\mathcal{U}}$ and $C_{\mathcal{M}}$, these criteria for each N have been estimated and ranked. Consider the optimal number of cluster gets minimum of these ranks:

$$N = \arg\min_N(\text{rank}(C_{\mathcal{U}}, N) + \text{rank}(C_{\mathcal{M}}, N))$$

where rank is the function that gives rank for each estimation for current N.

4 Distance functions for mixed-scale data

In this section, distance functions are described for different scale types — linear (10), ordinal (11), nominal (13), and mixed (12) and (14). The distance function for mixed-scale dataset is proposed below.

4.1 Distance function for linear-scale data

Consider the generalized distance function for a linear-scale dataset:

$$r(\mathbf{x}_i, \mathbf{x}_q) = \left(\left(|\mathbf{x}_i - \mathbf{x}_q|^p \right)^T \mathbf{S}^{-1} |\mathbf{x}_i - \mathbf{x}_q|^p \right)^{1/(2p)} \tag{10}$$

where p is the number; $\mathbf{S}$ is the symmetric nonnegative definite matrix (for example, identity matrix $\mathbf{I}$); and exponentiation is proceeded per component: $\mathbf{x}^p = [x_1^p, \ldots, x_n^p]^{\mathrm{T}}$. The Euclidean metric corresponds to this formula with $\mathbf{S} = \mathbf{I}$ and $p = 1$:

$$r(\mathbf{x}_i, \mathbf{x}_q) = \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{x}_q)^2)^{1/2}.$$

The Manhattan distance corresponds to this formula with $\mathbf{S} = \mathbf{I}$ and $p = 0.5$:

$$r(\mathbf{x}_i, \mathbf{x}_q) = \sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{x}_q|.$$

4.2 Distance for ordinal-scaled data

Define matrix functions $\mathbf{H}^{j+}$ and $\mathbf{H}^{j-}$ for projection the object set $\mathbf{X}$ on feature j where the scale $\mathbb{L}_j$ is ordinal. Each component of vectors $\mathbf{H}_i^{j+}$ and $\mathbf{H}_i^{j-}$ determine the order between feature j of object i and other objects:

$$(\mathbf{H}_i{}^{j+})_l = \begin{cases} 1 & \text{whenever } x_{ij} \succ x_{lj}; \\ 0 & \text{otherwise;} \end{cases}$$

$$(\mathbf{H}_i{}^{j-})_l = \begin{cases} 1 & \text{whenever } x_{lj} \succ x_{ij}; \\ 0 & \text{otherwise.} \end{cases}$$

Let the distance function pdist be given by:

$$\text{pdist}(x_{ij}, x_{qj}) = \frac{m - \left(\langle \mathbf{H}_i{}^{j+}, \mathbf{H}_q{}^{j+} \rangle + \langle \mathbf{H}_i{}^{j-}, \mathbf{H}_q{}^{j-} \rangle \right)}{m} \tag{11}$$

where m is the number of objects in the dataset.

Theorem 2. If $\mathbb{L}_j$ is a totally ordered set, then pdist is a metric.

Proof. At first, let us prove that the range of the function is in $[0;1]$. Let x_{ij} be less than or equal to x_{qj} : $x_{ij} \leqslant x_{qj}$. Then

$$\langle \mathbf{H}_i{}^{j+}, \mathbf{H}_q{}^{j+} \rangle = ||\mathbf{H}_i^{j+}||_2^2, \quad \langle \mathbf{H}_i{}^{j-}, \mathbf{H}_q{}^{j-} \rangle = ||\mathbf{H}_q^{j-}||_2^2;$$

$$\text{pdist}(x_{ij}, x_{qj}) = \frac{m - ||\mathbf{H}_i^{j+}||_2^2 - ||\mathbf{H}_q^{j-}||_2^2}{m}.$$

The maximum of the function is not more than 1. The function pdist gets minimum whenever $x_{ij} = x_{qj}$, $\text{pdist}(x_{ij}, x_{ij}) = 0$. The function is symmetric. Let us prove that the function satisfies the subadditivity condition for each $\mathbf{x}_w \in \mathbb{X}$:

$$\text{pdist}(x_{ij}, x_{qj}) \leqslant \text{pdist}(x_{ij}, x_{wj}) + \text{pdist}(x_{wj}, x_{qj}).$$

The proof contains 3 cases:

$$x_{ij} \leqslant x_{qj} \leqslant x_{wj}; \quad x_{wj} \geqslant x_{ij} \geqslant x_{qj}; \quad x_{ij} \leqslant x_{wj} \leqslant x_{ij}.$$

Consider the first case, other cases can be proved similarly:

$$\text{pdist}(x_{ij}, x_{wj}) + \text{pdist}(x_{qj}, x_{wj}) = \frac{2m - ||\mathbf{H}_i^{j+}||_2^2 - ||\mathbf{H}_q^{j+}||_2^2 - 2||\mathbf{H}_w^{j-}||_2^2}{m}$$

$$\geqslant \frac{2m - ||\mathbf{H}_i^{j+}||_2^2 - ||\mathbf{H}_q^{j+}||_2^2 - 2||\mathbf{H}_q^{j-}||_2^2}{m}$$

$$= \frac{2m - ||\mathbf{H}_i^{j+}||_2^2 - m + ||\mathbf{H}_q^{j-}||_2^2 - 2||\mathbf{H}_q^{j-}||_2^2}{m} = \frac{m - ||\mathbf{H}_i^{j+}||_2^2 - ||\mathbf{H}_q^{j-}||_2^2}{m} = \text{pdist}(x_{ij}, x_{qj}). \quad \blacksquare$$

4.3 The generalization of HEOM and HMOM distance functions

Supplement the HEOM [22] function for ordinal-scale datasets:

$$d_1(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{n} r(x_{ij}, x_{qj})^2 \right)^{1/2} \tag{12}$$

where

$$\left.\begin{aligned} r(x_{ij}, x_{qj}) &= \begin{cases} \text{overlap}(x_{ij}, x_{qj}) & \text{whenever } \mathbb{L}_j \text{ is a nominal scale;} \\ \text{pdist}(x_{ij}, x_{qj}) & \text{whenever } \mathbb{L}_j \text{ is an ordinal scale;} \\ \text{diff}(x_{ij}, x_{qj}) & \text{otherwise;} \end{cases} \\ \text{overlap}(x_{ij}, x_{qj}) &= \begin{cases} 1 & \text{whenever } x_{ij} \neq x_{qj}; \\ 0 & \text{otherwise;} \end{cases} \\ \text{diff}(x_{ij}, x_{qj}) &= \frac{|x_{ij} - x_{qj}|}{\max\limits_{\mathbb{L}_j} - \min\limits_{\mathbb{L}_j}}, \end{aligned}\right\} \tag{13}$$

the function $\text{diff}(x_{ij}, x_{qj})$ is determined by normalized difference between two values of feature j.

The range of the resulting function d is less than or equal to the square root of the feature number: $d(\mathbf{x}_i, \mathbf{x}_j) \leqslant \sqrt{n}$.

The difference between HEOM and HMOM modifications is only in lack of exponentiation:

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{n} r(x_{ij}, x_{qj}). \tag{14}$$

5 Panel matrix recovery procedure

5.1 Clustering algorithm $\mathfrak{c}$

Let use a modification of k-means [26] algorithm as the clustering algorithm $\mathfrak{c}$. This algorithm is iterative. At first, select N cluster centroids $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ randomly. Each iteration assign each object $\mathbf{x}_i$ from dataset $\mathbf{X}^t$ to the closest cluster in the sense of the distance function d:

$$\text{cluster}(\mathbf{x_i}) = \arg\min_{k \in \{1,\ldots,N\}} d(\mathbf{x}_i, \boldsymbol{\mu}_k)$$

where $\text{cluster}(\mathbf{x})$ is the function that returns a cluster index for each object $\mathbf{x}$. After that, recalculate cluster centroids:

$$\mu_{kj} = \text{avg}\{x_{ij}, \text{cluster}(\mathbf{x}_i) = k\}.$$

Use avg function (6) such that corresponds to scale types instead of arithmetic mean recommended in the k-means algorithm:

$$\text{avg}\{\mathbf{x}_{i_1 j}, \ldots, \mathbf{x}_{i_p j}\} = \begin{cases} \text{mean}\{\mathbf{x}_{i_1 j}, \ldots, \mathbf{x}_{i_p j}\} & \text{whenever } \mathbb{L}_j \text{ is a linear scale;} \\ \text{median}\{\mathbf{x}_{i_1 j}, \ldots, \mathbf{x}_{i_p j}\} & \text{whenever } \mathbb{L}_j \text{ is an ordinal scale;} \\ \text{mode}\{\mathbf{x}_{i_1 j}, \ldots, \mathbf{x}_{i_p j}\} & \text{whenever } \mathbb{L}_j \text{ is a nominal scale.} \end{cases}$$

5.2 Bijection recovery algorithm $\mathfrak{m}$

In this section, two methods of the function φ (3) finding are considered: the reducing the problem to the transport problem and the genetic algorithm.

Let state the problem of finding φ as multidimensional assignment problem [7]. Construct $|\mathcal{T}|$-partite hypergraph $\langle V, E \rangle$, $V = V^1 \cup \cdots \cup V^{|\mathcal{T}|}$ where $\mathcal{T}$ is the set of years. The vertices of each partite sets V^t correspond to the set of cluster centroids $\mathbf{M}^t$ for year t. The hyperedges of the hypergraph correspond to the all subsets of cluster centroids that contain $|\mathcal{T}|$ cluster centroids and correspond to the condition that each hyperedge $e \in E$ contains only one cluster centroid for each year. Let the weight of each hyperedge be given by:

$$w_e = \sum_{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in e} \rho(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2).$$

It is required to find a maximal set of hyperedges where each pair of this set does not intersect and where the sum of the hyperedge weights is minimal.

5.3 Reducing the φ finding problem to the transport problem

Consider a two-dimensional assignment problem, where it is required to find the bijection between two sets of objects. This problem can be stated as the min-cost max flow problem by constructing a transport directed graph [27]. The vertices of this graph correspond to the cluster centroids $\boldsymbol{M}^t$ with capacity equal to one and edge weights equal to the distance between cluster centroids: $\rho(\boldsymbol{\mu}_i^{t_1}, \boldsymbol{\mu}_j^{t_2})$. After reducing the problem, it is required to find the maximal flow with minimal edge weight sum, which is called the cost of the flow. In the considered case, one can construct a hypergraph $\langle V, E \rangle$ instead of the directed graph whose hyperedge configuration was described above.

In order to find the maximal flow of minimal cost of the hypergraph $\langle V, E \rangle$, let transform the hypergraph into a directed graph $\langle V', E' \rangle$ and use common algorithms for directed graphs.

There are some heuristic algorithms of hypergraph to directed graph transformation that can be used for this case [28, 29].

5.4 Genetic algorithm

As an alternative method of finding the function φ, let use the genetic algorithm [8]. Each solution of the problem is represented by a hypergraph with N hyperedges such that each pair of hyperedge does not intersect and each hyperedge contains only one cluster centroid for each year. Let $\mathbf{S}^{qk}$ be a matrix for the solution k of the generation q. The entry S_{ij}^{qk} is the index number of cluster of the year j in the hyperedge i:

$$S_{ij}^{qk} = \text{whenever } \boldsymbol{\mu}_l^j \in e_i$$

where $e_i \in E$. The starting population $\mathbf{S}^1$ is generated randomly, its cardinality s_1 is a structural parameter. Each new generation is generated from the older one by application the special procedures: mutation, crossover, and selection.

As the crossover of the generation q, the following procedure has been used. Select two solutions $\mathbf{S}^{qk_1}$ and $\mathbf{S}^{qk_2}$ from this generation randomly. Also, select a row l_1 from the first matrix and a row l_2 from the second matrix, the number of columns to modify col, and a set of column indexes $\{c_{\text{perm}_{(1)}}, \ldots, c_{\text{perm}_{(\text{col})}}\}$, where perm is a random permutation. For each column c_i in $\{c_{\text{perm}_{(1)}}, \ldots, c_{\text{perm}_{(\text{col})}}\}$ and for both matrices, proceed the permutation given by $\mathbf{S}_{l_1 c_i}^{qk_1} \leftrightarrow \mathbf{S}_{l_2 c_i}^{qk_2}$.

After crossover procedure, mutations. Select a solution $\mathbf{S}^{qk}$ and a column c randomly. After that, proceed random permutation on all the elements of column c. Such procedure helps one to avoid stopping algorithm in local extrema. After mutations and crossovers, select the best solution generation $\mathbf{S}^{q+1}$ in the sense of the distance function ρ. The number of mutations per generation f_{mutation}, the number of crossovers per generation $f_{\text{crossover}}$, and the generation cardinalities s_q are the structural parameters of the algorithm. The algorithm stops whenever the generation satisfies the stopping criterion $C_{\mathcal{F}}$. In this paper, stopping criterion is used:

$$C_{\mathcal{F}} = (K_{\text{av}} > \hat{K}_{\text{av}}) \text{ or } (K_{\text{av}} \text{ does not change after few iterations})$$

where

$$K_{\text{av}} = \text{mean}_{t_1,t_2 \in \mathcal{T}, t_1 \neq t_2}(\text{KendallTau}(\mathbf{S}_{1,\ldots,N,t_1}^{q(1)}, \mathbf{S}_{1,\ldots,N,t_2}^{q(1)})),$$

$\mathbf{S}^{q(1)}$ is the best solution of the current generation in the sense of ρ, $\mathbf{S}_{1,\ldots,N,t}^{qi}$ is the column t of the matrix $\mathbf{S}^{qi}$. $\hat{K}_{\text{av}}$ is a structural parameter, which represents required average Kendall correlation coefficient in the panel matrix $\mathbf{Z}$.

5.5 Defining hyperedge weight

Use the sum of the generalized distances (12) and (14) between cluster centroids as the hyperedge weights:

$$\rho_1(\mathbf{x}_i, \mathbf{x}_j) = \Big(\frac{\sum_{k=1}^n d(x_{ik}, x_{jk})^2}{n} + \text{pdist}^2(y_i, y_j) \cdot \text{coef} \Big)^{1/2}; \tag{15}$$

$$\rho_2(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{k=1}^n d(x_{ik}, x_{jk})}{n} + \text{pdist}(y_i, y_j) \cdot \text{coef} \tag{16}$$

where coef is the parameter which regulates the balance between priority of the stability criterion (5) and the clustering criterion (4). Whenever coef = 1, these criteria priorities are equal. Let use (15) in the experiment with the generalized HEOM metric and (16) in the experiment with the generalized HMOM metric.

5.6 Complexity analysis of the algorithm

The clustering algorithm complexity can be bounded to $O(Nnm \cdot \text{iter})$ where iter is the number of iterations of the clustering algorithm.

The complexity of one crossover series can be bounded to $O(f_{\text{crossover}}|\mathcal{T}|)$. The complexity of a mutation series is $O(f_{\text{mutation}}N)$; so, the naive estimation of the genetic algorithm iteration is $O(f_{\text{crossover}}|\mathcal{T}| + f_{\text{mutation}}N)$.

6 The ranking model recovery

In this section, the methods of the ranking model recovery used in this paper are described. Consider three ranking algorithms: the ordinal classification algorithm using partially ordered feature sets [18] and rankSVM [30], the algorithm based on the SVM [19], and an algorithm based on the method of least squares in order to compare the results of the ranking model recovery.

6.1 The ordinal classification algorithm using partially ordered feature sets

In this subsection, suppose that the class of the object is also a feature with number 0: $\mathbb{Y} = \mathbb{L}_0, x_{i0} = y_i, \mathbf{x}_i \in \mathbf{X}$. For each feature $\mathbb{L}_q$, construct a matrix $\mathbf{U}_q$ which determines the order of the feature q:

$$U_q(i,j) = \begin{cases} 1 & \text{if } x_{iq} \prec x_{jq}; \\ 0 & \text{otherwise}. \end{cases}$$

Estimate the matrix $\boldsymbol{\psi}$ using feature matrices $\mathbf{U}_q, q \in \{0, \ldots, m\}$. This matrix $\boldsymbol{\psi}$ is called a pairwise dominance matrix:

$$\hat{\psi_{ij}} = \sum_{k=1}^{n} w_k U_k(i,j);$$

$$\mathbf{w} = \arg\min_{\mathbf{w}} \sum_{i=1}^{m}\sum_{k=1}^{m} \left(U_0(i,k) - \sum_{j=1}^{n} w_j U_j(i,k) \right)^2$$

where $\mathbf{w}$ is the weight vector for feature matrices.

After that, estimate the class $\hat{y}$ of the object using the pairwise dominance matrix:

$$\hat{y} = f(\hat{\psi}, \lambda), \quad \lambda = \arg\min_{\lambda} ||y - \hat{y}||_2.$$

In this paper, a logistic regression is used for $\mathbf{w}$ and λ estimations. Also, propose that $\lambda_i = \lambda_j$ whenever $y_i = y_j$.

6.2 The RankSVM algorithm

This algorithm is a generalization of the classification algorithm based on SVM [19]. The optimization problem for this algorithm is given by

$$||\mathbf{w}||_2 + C \sum_{i,j} \xi_{ij} \to \min,$$

$$\text{for each } \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}, y_y > y_j : K(\mathbf{w}, \mathbf{x}_i) \geqslant K(\mathbf{w}, \mathbf{x}_j) + 1 - \xi_{ij}$$

where

$$K : \mathbb{R}^n \times \mathbb{X} \to \mathbb{R} \tag{17}$$

is the kernel function, commonly the dot product; ξ_{ij} and C are the parameters. This optimization problem can be reduced to the classification SVM optimization problem [30] and solved by standard methods [19].

The most interesting feature of this algorithm is the use of different kernel functions K instead of dot product. This modifies original object space and makes it more similar to linearly separable space. In this paper, the following kernel functions have been used:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j^{\mathrm{T}}; \tag{18}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left(\frac{1}{n}\mathbf{x}_i \cdot \mathbf{x}_j^{\mathrm{T}}\right)^3; \tag{19}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{n}|\mathbf{x}_i - \mathbf{x}_j|^2\right); \tag{20}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh\left(\frac{1}{n}\mathbf{x}_i \cdot \mathbf{x}_j^{\mathrm{T}}\right). \tag{21}$$

6.3 The algorithm based on least squares method

Use this algorithm as a basic ranking algorithm. The main idea of this algorithm is in finding coefficients $\alpha_1, \ldots, \alpha_n$, which solve the optimization task:

$$\Delta = \sum_{i=1}^{m} \left\| y_i - \sum_{j=1}^{n} \alpha_j x_{ij} \right\|_2 \to \min.$$

The resulting function is given by

$$f(\mathbf{x}) = \begin{cases} \text{round}\left(\sum_{j=1}^{n} \alpha_j x_i j\right) & \text{whenever round}\left(\sum_{j=1}^{n} \alpha_j x_{ij}\right) \in \{1,2,3,4,5\}; \\ 5 & \text{whenever round}\left(\sum_{j=1}^{n} \alpha_j x_{ij}\right) > 5; \\ 1 & \text{whenever round}\left(\sum_{j=1}^{n} \alpha_j x_{ij}\right) < 1. \end{cases}$$

6.4 Transforming ordinal features into linear features

In order to use the information gathered from ordinal features, the following approach has been used [31]. Each ordinal feature with scale $\mathbb{L}_j$ is proposed to match with some latent linear feature with scale $\mathbb{L}_j^*$, which can be recovered by the following rule:

$$x_{ij} = l_{ju} \text{ whenever } l^*_{ju-1} \leqslant x^*_{ij} \leqslant l_{ju}$$

where l_{ju} is the u value of the set of values of $\mathbb{L}_j$ sorted ascendingly; x^*_{ij} is the value of latent variable; l^*_{ju} is the threshold:

$$l_{ju} = \Psi^{-1}(F_j(l_{ju})), \quad F_j(l) = \sum_{\mathbf{x}_i \in \mathbf{X}, x_{ij} \prec l} \frac{1}{m},$$

$$l_{j0} = -\infty, \quad l_{j|\mathbb{L}_j|} = \infty$$

with Ψ^{-1} being the inverse normal distribution. This transformation matches the ordinal feature with some real-valued intervals. Use the upper limit of the intervals as a representer of the latent

linear feature, i. e., $x^*_{ij} = l_{ju}$ whenever $l^*_{ju-1} \leqslant x^*_{ij} \leqslant l_{ju}$. Let the value corresponding to the largest value of the ordinal feature be $x^*_{ij} = l_{j|\mathbb{L}_j|-1} + \text{mean}(\{l_{ju} - l_{ju-1}, u \in \{1, \ldots, |\mathbb{L}_j| - 1\})$.

7 Computational experiment

In this section, the results of the experiment are presented and conclusions on the applicability of the proposed algorithm to the considered problem are drawn. The main goal of the present experiment is to confirm or deny the efficiency of the described panel matrix recovery method and recover the ranking model in the most efficient way. The dataset [20] contains a table with 284 student assessments. Each assessment contains 7 features, the class of the student, and the year of the interview. The source of the computational experiment is available at [32].

For the experiment, the following software was used:

- GNU Octave v.3.8.1;
- SVM$^{\text{light}}$ v.6.02.;
- batch high throughput multidimensional scaline for MATLAB/GNU Octave programming language; and
- Python v.2.7. with NumPy and scikit-learn packages.

7.1 Panel matrix recovery

In order to handle with missing values, k-nearest neighbors algorithm has been used for missing values imputation [33]. $k = 3$ was chosen using cross-validation. The optimal number of clusters N has been estimated by solving the optimization problem (9) and the result $N = 20$ has been got. The results of clustering for year 2007 are shown in Fig. 4. The coordinates of the objects were received by projection the data $\mathbf{X}^t$ onto two-dimensional space $\{\xi_1, \xi_2\}$ using High-Throughput Multidimensional Scaling [34] method. The colors of the plot correspond to different cluster indexes.

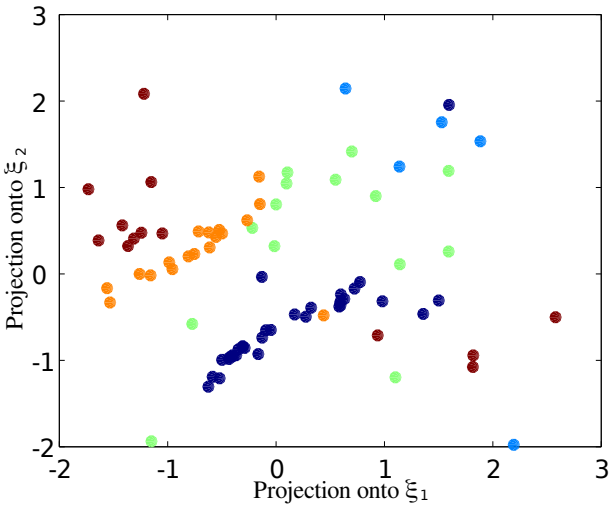

Figure 4 The result of clustering for year 2007

In order to reduce the randomnicity in the experiment, 10 tests have been proceeded and the results have been averaged. The parameter coef (15) was set to 1. The cardinality of the starting population s_1 was set to $N \cdot |\mathcal{T}|$. For each generation $\mathbf{S}^q$, $|\mathcal{T}| \cdot s^q$ mutations and $\binom{s^q}{2} = s^q(s^q - 1)/2$ crossovers have been proceeded. Such parameter values give an availability to crossover all the pairs of solutions and to mutate each column of each hypergraph matrix. The

cardinality of all the generations remained stable: $s^{q+1} = s^q$. The required average Kendall coefficient $\hat{K}_{\text{av}}$ was set to 0.85.

The results of the panel matrix recovery have been estimated by the Kendall correlation coefficient (7). The results of the Kendall correlation for the experiments with HEOM and HMOM metrics are represented in Tables 2 and 3. The computational experiment shows that the proposed algorithm of the panel matrix recovery gives good results on the considered dataset.

Table 2 Kendall coefficient for panel matrix **Z** recovery with HEOM metric

Year	2006	2007	2008	2009
2006	1	0.85629	0.80154	0.85270
2007	0.85629	1	0.84301	0.85728
2008	0.80154	0.84301	1	0.84731
2009	0.85270	0.85728	0.84731	1

Table 3 Kendall coefficient for panel matrix **Z** recovery with HMOM metric

Year	2006	2007	2008	2009
2006	1	0.87714	0.76905	0.77979
2007	0.87714	1	0.82962	0.80129
2008	0.76905	0.82962	1	0.82266
2009	0.77979	0.80129	0.82266	1

The mean of pairwise Kendall coefficients for the panel recovery with HMOM is 0.81326 while for HEOM, this value is 0.84302. Therefore, HEOM metric is quite more efficient for the considered purpose. As one can see, the panel matrix recovery gives rather stable results for all the years.

7.2 The ranking model

The difference between the real class y of an object $\mathbf{x}$ and the recovered class $\hat{y}$ has been used as the error function Q. The algorithms were tested using "Leave one out" method. Different kernel functions (17) have been used during the RankSVM algorithm testing.

The results of the experiment are shown in Table 4.

The RankSVM algorithm showed the best result and was selected as the ranking model recovery algorithm. Another good result was received from the algorithm based on pairwise-dominating matrix.

7.3 Computation for the simulated data

Investigate the performance of the proposed algorithm. Conduct two series of the experiments: the series with adding noise into object features and adding noise into object classes.

Table 4 Results of the ranking model recovery

Year	2006	2007	2008	2009	Mean value
LS-algrotihm	0.7	0.57	0.68	0.62	0.64
Pairwise-dominating matrix	1.2176	1.1412	1.2647	1.2235	1.2118
RankSVM, Eq. (18)	0.55	0.52	0.62	0.60	**0.58**
RankSVM, Eq. (19)	1,2741	0.98	1.3451	1.1667	1.1914
RankSVM, Eq. (20)	0.7511	0.5413	0.7285	0.7501	0.6927
RankSVM, Eq. (21)	1.2741	0.98	1.3451	1.1667	1.1914

During the experiment with adding noise into features, change each object feature value randomly with probability from 10% to 50%.

During the experiment with adding noise in classes, replace the class of each object by constant for each year. In these experiments, HEOM metric has been used.

The results of the experiments are shown in Figs. 5 and 6.

Figure 5 shows the mean of Kendall correlation coefficient values for each pairs of years. The genetic algorithm uses the combination of two criteria for selecting optimal solution. After adding noise into objects' features, the algorithm tries to optimize the matching of classes for different years. The experiment with adding noise into object classes shows the opposite case — the dataset lost the uniformity of classes per years and, therefore, the average error did not increase so dramatically as in the first experiment.

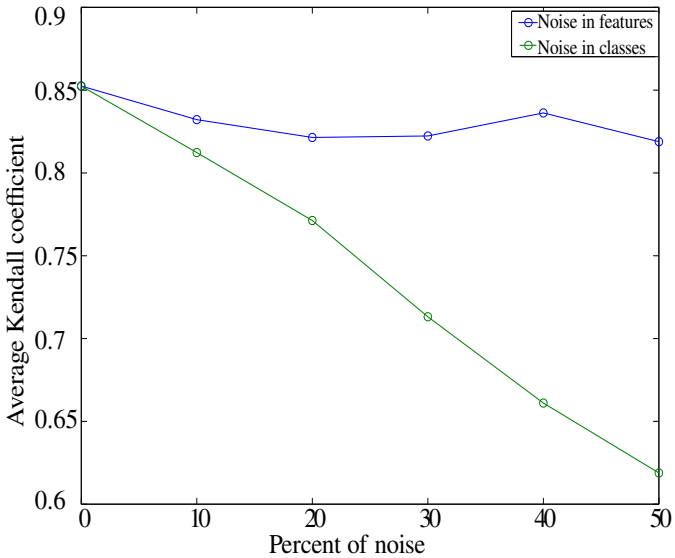

Figure 5 Average Kendall coefficient

In order to estimate the quality of the ranking model recovery in datasets with noise, the ranking model recovery has been tested on the simulated datasets generated in the first experiment series.

The result of the ranking model recovery is shown in Fig. 7. As one can see, the RankSVM algorithm and pairwise-dominating algorithm give the similar results.

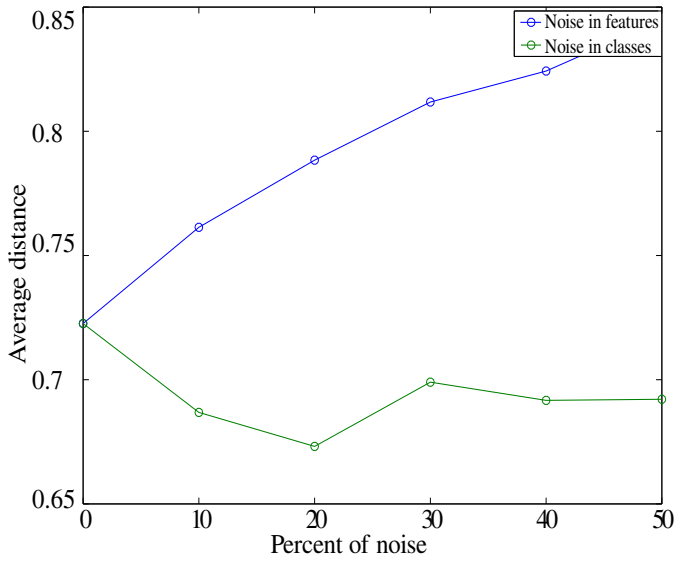

Figure 6 Average distance between cluster centroids that were mapped by the bijection

The error of the least squares-based algorithm increases dramatically; therefore, it is better not to use it if the dataset contains significant amount of noise.

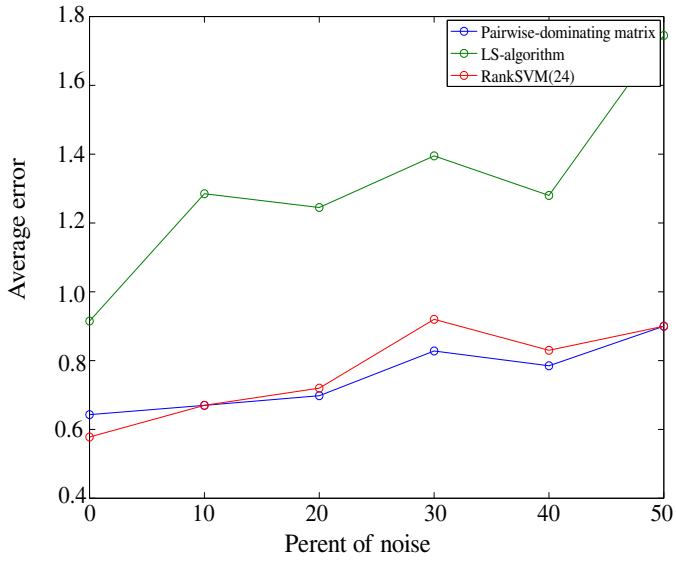

Figure 7 The results of rank model recovery on simulated data

8 Concluding remarks

In this paper, the method of the panel matrix recovery has been proposed. The heuristic method of calculating optimal number of clusters has been suggested for clustering objects per year.

Two algorithms have been considered to construct a bijection between clusters of different years based on reducing this problem to multidimensional assignment problem — the ge-

netic algorithm and the algorithm based on the reducing the problem to the transport problem.

The experiment for the panel matrix and ranking model recovery using genetic algorithm was proceeded. Two metric functions were compared. The HEOM metric showed the best result. The experiment showed that the panel matrix was stable in the sense of ranking model stability. The best result of ranking model recovery was shown by the RankSVM algorithm.

9 Acknowledgments

The author would like to thank Dr. Vadim Strijov for the formal problem statement, useful comments, suggestions, and the attention drawn to the present work.

References

[1] Davies, A., and K. Lahiri. 1995. A new framework for testing rationality and measuring aggregate shocks using panel data. *J. Econometrics* 68(1):205–227.

[2] Capponi, A., and H. de Waard. 2004. A polynomial time algorithm for the multidimensional assignment problem in multiple sensor environments. *7th Conference (International) on Information Fusion*. 1150–1157.

[3] Pardalos, P., and L. Pitsoulis. 2001. *Nonlinear assignment problems: Algorithms and applications (combinatorial optimization)*. Springer. 303 p.

[4] Aronson, J. E. 1986. The multiperiod assignment problem: A multicommodity network flow model and specialized branch and bound algorithm. *Eur. J. Oper. Res.* 23(3):367–381.

[5] Fréville, A. 2004. The multidimensional 0–1 knapsack problem: An overview. *Eur. J. Oper. Res.* 155(1):1–21. doi: http://dx.doi.org/10.1016/S0377-2217(03)00274-1

[6] Walteros, J. L., C. Vogiatzis, E. L. Pasiliao, and P. M. Pardalos. 2014. Integer programming models for the multidimensional assignment problem with star costs. *Eur. J. Oper. Res.* 235(3):553–568. doi: http://dx.doi.org/10.1016/j.ejor.2013.10.048

[7] Kuroki, Y., and T. Matsui. 2009. An approximation algorithm for multidimensional assignment problem minimizing the sum of squared errors. *Discrete Appl. Math.* 157:2124–2135.

[8] Sahu, A., and R. Tapadar. 2007. Solving the assignment problem using genetic algorithm and simulated annealing. *IAENG Int. J. Appl. Math.* 36(1). Available at: http://www.iaeng.org/IJAM/issues_v36/issue_1/IJAM_36_1_7.pdf (accessed January 9, 2016).

[9] Pistorius, J., and M. Minoux. 2003. An improved direct labeling method for the max-flow min-cut computation in large hypergraphs and applications. *Int. Trans. Oper. Res.* 10(1):1–11.

[10] Cooke, D. J., and H. E. Bez. 1984. *Computer mathematics*. 1st ed. Cambridge University Press. 408 p.

[11] Johannes, F., and H. Eyke. 2011. *Preference learning: An introduction*. Springer. 454 p.

[12] Albadvi, A. 2004. Formulating national information technology strategies: A preference ranking model using PROMETHEE method. *Eur. J. Oper. Res.* 153(2):290–296. doi: http://dx.doi.org/10.1016/S0377-2217(03)00151-6

[13] Siskos, Y., N. F. Matsatsinis, and G. Baourakis. 2001. Multicriteria analysis in agricultural marketing: The case of French olive oil market. *Eur. J. Oper. Res.* 130(2):315–331. doi: http://dx.doi.org/10.1016/S0377-2217(00)00043-6

[14] Mladineo, N., J. Margeta, J. P. Brans, and B. Mareschal. 1987. Multicriteria ranking of alternative locations for small scale hydro plants. *Eur. J. Oper. Res.* 31(2):215–222. doi: http://dx.doi.org/10.1016/0377-2217(87)90025-7

[15] Strijov, V. V. 2011. Utochnenie ekspertnykh otsenok, vystavlennykh v rangovykh shkalakh, s pomoshch'yu izmeryaemykh dannykh [Clarification of expert estimations in rank scale using measured data]. *Zavodskaya laboratoriya. Diagnostika materialov* [Factory Laboratory. Material Diagnostics] 77(7):72–78. (In Russian.)

[16] Medvednikova, M. M. 2012. Ispol'zovanie metoda glavnykh komponent pri postroenii integral'nykhx indikatorov [Using principal component analysis in construction of integral indicators]. *Machine Learning Data Anal.* 1(3):292–304. (In Russian.)

[17] Medvednikova, M. M., V. V. Strijov, and M. P. Kuznetsov. 2012. Algoritm mnogoklassovoy monotonnoy Pareto-klassifikatsii s vyborom priznakov [An algorithm of multiclass monotonic Pareto-classification with feature selection]. *Proceedings of the Tula State University, Natural Sciences* 3:132–141. (In Russian.)

[18] Kuznetsov, M. P., and V. V. Strijov. 2014. Methods of expert estimations concordance for integral quality estimation. *Expert Syst. Appl.* 41(4, Pt. 2):1551–2110. doi: http://dx.doi.org/10.1016/j.eswa.2013.08.095

[19] Vorontsov, K. V. Support vector machine lectures. Available at: http://www.ccas.ru/voron/download/SVM.pdf (accessed July 22, 2014). (In Russian.)

[20] The original dataset. Available at: http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group074/Bakhteev014UniversityRanking/data/data.csv?format=raw (accessed August 27, 2014).

[21] Strijov, V. V. 2006. Utochnenie ekspertnykh otsenok s pomoshch'yu izmeryaemykh dannykh [Clarification of expert estimations using measured data]. *Zavodskaya laboratoriya. Diagnostika materialov* [Factory Laboratory. Material Diagnostics] 72(7):59–64. (In Russian.)

[22] Wilson, D. R., and T. R. Martinez. 1997. Improved heterogeneous distance functions. *J. Artif. Intell. Res.* 6:1–34.

[23] Batista, G. E. A. P. A., and D. F. Silva. 2009. How k-nearest neighbor parameters affect its performance. *Argentine Symposium on Artificial Intelligence Proceedings*. 1–12.

[24] Walesiak, M. 1999. Distance measure for ordinal data. *Argum. Oecon.* 2(8):167–173.

[25] Prokhorov, A. V. (originator) Kendall coefficient of rank correlation. *Encyclopedia of mathematics*. Available at: http://www.encyclopediaofmath.org/index.php?title =Kendall_coefficient_of_rank_correlation&oldid=13189 (accessed August 27, 2014).

[26] Steinhaus, H. 1956. Sur la division des corps materiels en parties. *Bull. Acad. Pol. Sci.* 4:801–804.

[27] Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein. 2009. *Introduction to algorithms*. 3rd ed. MIT Press. 1312 p.

[28] Agarwal, S., K. Branson, and S. Belongie. 2006. Higher order learning with graphs. *23rd Conference (International) on Machine Learning Proceedings*. 17–24. doi: http://dx.doi.org/10.1145/1143844.1143847

[29] Pu, L., and B. Faltings. 2012. Hypergraph learning with hyperedge expansion. *European Conference on Machine Learning and Knowledge Discovery in Databases Proceedings*. 1:410–425. doi: http://dx.doi.org/10.1007/978-3-642-33460-3_32

[30] Joachims, T. 2002. Optimizing search engines using clickthrough data. *8th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining Proceedings*. 133–142. doi: http://dx.doi.org/10.1145/775047.775067

[31] Winship, C., and R. Mare. 1984. Regression models with ordinal variables. *Am. Sociol. Rev.* 49(4):512–525. doi: http://dx.doi.org/10.2307/2095465

[32] Algorithms of machine learning. Available at: http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group074/Bakhteev014UniversityRanking/code/ (accessed August 27, 2014).

[33] Batista, G. E. A. P. A., and M. C. Monard. 2003. An analysis of four missing data treatment methods for supervised learning. *Appl. Artif. Intell.* 17(5-6):519–533. doi: http://dx.doi.org/10.1080/713827181

[34] Strickert, M., S. Teichmann, N. Sreenivasulu, and U. Seiffert. 2005. High-Throughput Multi-dimensional Scaling (HiT-MDS) for cDNA-Array expression data. *15th Conference (International) on Artificial Neural Networks: Biological Inspirations Proceedings*. 1:625–633. doi: http://dx.doi.org/10.1007/11550822_97

Received December 20, 2015

Восстановление панельной матрицы и ранжирующей модели по метризованной выборке в разнородных шкалах

О. Ю. Бахтеев

bakhteev@phystech.edu

Московский физико-технический институт, Россия, г. Долгопрудный, Институтский пер., 9

Работа посвящена восстановлению ежегодных изменений рейтингов студентов при собеседовании в учебный центр. Рассматривается выборка, состоящая из экспертных оценок студентов, проходивших собеседование в учебный центр в течение нескольких лет и итоговых рейтингов студентов. Шкалы экспертных оценок меняются из года в год, но шкала рейтингов остается неизменной. Требуется восстановить ранжирующую модель, не зависящую от времени. Задача сводится к восстановлению панельной матрицы (т. е. матрицы объект–признак–год), ставящей во взаимное соответствие некоторого студента (или усредненный «портрет» студента) и его предполагаемую оценку на собеседованиях за каждый год, и исследованию ранжирующей модели, полученной на основе этой матрицы, а также анализу ее устойчивости на протяжении нескольких лет. Предлагается метод восстановления панельной матрицы, основанный на решении многомерной задачи о назначениях. В качестве метода восстановления ранжирующей модели используется алгоритм многоклассовой классификации с отношением полного порядка на классах.

Ключевые слова: *рейтинги; экспертные оценки; кластеризация; смешанные шкалы*

DOI: 10.21469/22233792.1.14.05

Литература

[1] *Davies A., Lahiri K.* A new framework for testing rationality and measuring aggregate shocks using panel data // J. Econometrics, 1995. Vol. 68. No. 1. P. 205–227.

[2] *Capponi A., de Waard H.* A polynomial time algorithm for the multidimensional assignment problem in multiple sensor environments // 7th Conference (International) on Information Fusion, 2004. P. 1150–1157.

[3] *Pardalos P., Pitsoulis L.* Nonlinear assignment problems: Algorithms and applications (combinatorial optimization). — Springer, 2001. 303 p.

[4] *Aronson J. E.* The multiperiod assignment problem: A multicommodity network flow model and specialized branch and bound algorithm // Eur. J. Oper. Res., 1986. Vol. 23. No. 3. P. 367–381.

[5] *Fréville A.* The multidimensional 0—1 knapsack problem: An overview // Eur. J. Oper. Res., 2004. Vol. 155. Iss. 1. P. 1—21. doi: http://dx.doi.org/10.1016/S0377-2217(03)00274-1

[6] *Walteros J. L., Vogiatzis C., Pasiliao E. L., Pardalos P. M.* Integer programming models for the multidimensional assignment problem with star costs // Eur. J. Oper. Res., 2014. Vol. 235. No. 3. P. 553–568. doi: http://dx.doi.org/10.1016/j.ejor.2013.10.048

[7] *Kuroki Y., Matsui T.* An approximation algorithm for multidimensional assignment problem minimizing the sum of squared errors // Discrete Appl. Math., 2009. Vol. 157. P. 2124–2135.

[8] *Sahu A., Tapadar R.* Solving the assignment problem using genetic algorithm and simulated annealing // IAENG Int. J. Appl. Math., 2007. Vol. 36. No. 1. http://www.iaeng.org/IJAM/issues_v36/issue_1/IJAM_36_1_7.pdf.

[9] *Pistorius J., Minoux M.* An improved direct labeling method for the max-flow min-cut computation in large hypergraphs and applications // Int. Trans. Oper. Res., 2003. Vol. 10. No. 1. P. 1–11.

[10] *Cooke D. J., Bez H. E.* Computer mathematics. — 1st ed. — Cambridge University Press, 1984. 408 p.

[11] *Johannes F., Eyke H.* Preference learning: An introduction. — Springer, 2011. 454 p.

[12] *Albadvi A.* Formulating national information technology strategies: A preference ranking model using PROMETHEE method // Eur. J. Oper. Res., 2004. Vol. 153. No. 2. P. 290–296. doi: http://dx.doi.org/10.1016/S0377-2217(03)00151-6

[13] *Siskos Y., Matsatsinis N. F., Baourakis G.* Multicriteria analysis in agricultural marketing: The case of French olive oil market // Eur. J. Oper. Res., 2001. Vol. 130. No. 2. P. 315–331. doi: http://dx.doi.org/10.1016/S0377-2217(00)00043-6

[14] *Mladineo N., Margeta J., Brans J. P., B. Mareschal.* Multicriteria ranking of alternative locations for small scale hydro plants // Eur. J. Oper. Res., 1987. Vol. 31. No. 2. P. 215–222. doi: http://dx.doi.org/10.1016/0377-2217(87)90025-7

[15] *Стрижов В. В.* Уточнение экспертных оценок, выставленных в ранговых шкалах, с помощью измеряемых данных // Заводская лаборатория. Диагностика материалов, 2011. Т. 77. № 7. С. 72–78.

[16] *Медведникова М. М.* Использование метода главных компонент при построении интегральных индикаторов // Машинное обучение и анализ данных, 2012. Т. 3. С. 292–304.

[17] *Медведникова М. М., Стрижов В. В., Кузнецов М. П.* Алгоритм многоклассовой монотонной Парето-классификации с выбором признаков // Известия Тульского гос. ун-та. Естественные науки, 2012. Т. 3. С. 132–141.

[18] *Kuznetsov M. P., Strijov V. V.* Methods of expert estimations concordance for integral quality estimation // Expert Syst. Appl., 2014. Vol. 41. Iss. 4. Pt. 2. P. 1551–2110. doi: http://dx.doi.org/10.1016/j.eswa.2013.08.095

[19] *Воронцов К. В.* Лекции по методу опорных векторов. http://www.ccas.ru/voron/download/SVM.pdf.

[20] http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group074/Bakhteev014UniversityRanking/data/data.csv?format=raw.

[21] *Стрижов В. В.* Уточнение экспертных оценок с помощью измеряемых данных // Заводская лаборатория. Диагностика материалов, 2006. Т. 72. № 7. С. 59–64.

[22] *Wilson D. R., Martinez T. R.* Improved heterogeneous distance functions // J. Artif. Intell. Res., 1997. Vol. 6. P. 1–34.

[23] *Batista G. E. A. P. A., Silva D. F.* How k-nearest neighbor parameters affect its performance // Argentine Symposium on Artificial Intelligence Proceedings, 2009. P. 1–12.

[24] *Walesiak M.* Distance measure for ordinal data // Argum. Oecon., 1999. Vol. 2. No. 8. P. 167–173.

[25] *Prokhorov A. V.* (originator). Kendall coefficient of rank correlation // Encyclopedia of mathematics. http://www.encyclopediaofmath.org/index.php?title=Kendall_coefficient_of_rank_correlation&oldid=13189.

[26] *Steinhaus H.* Sur la division des corps materiels en parties // Bull. Acad. Polon. Sci., 1956. Vol. 4. P. 801–804.

[27] *Cormen T. H., Leiserson C. E., Rivest R. L., Stein C.* Introduction to algorithms. — 3rd. ed. — MIT Press, 2009. 1312 p.

[28] *Agarwal S., Branson K., Belongie S.* Higher order learning with graphs // 23rd Conference (International) on Machine Learning Proceedings, 2006. P. 1–24. doi: http://dx.doi.org/10.1145/1143844.1143847

[29] *Pu L., Faltings B.* Hypergraph learning with hyperedge expansion // European Conference on Machine Learning and Knowledge Discovery in Databases Proceedings, 2012. Vol. 1. P. 410–425. doi: http://dx.doi.org/10.1007/978-3-642-33460-3_32

[30] *Joachims T.* Optimizing search engines using clickthrough data // 8th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining Proceedings, 2002. P. 133–142. doi: http://dx.doi.org/10.1145/775047.775067

[31] *Winship C., Mare R.* Regression models with ordinal variables // Am. Sociol. Rev., 1984. Vol. 49. No. 4. P. 512–525. doi: http://dx.doi.org/10.2307/2095465

[32] Algorithms of machine learning. http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group074/Bakhteev014UniversityRanking/code/.

[33] *Batista G. E. A. P. A., Monard C. M.* An analysis of four missing data treatment methods for supervised learning // Appl. Artif. Intell., 2003. Vol. 17. Iss. 5-6. P. 519–533. doi: http://dx.doi.org/10.1080/713827181

[34] *Strickert M., Teichmann S., Sreenivasulu N., Seiffert U.* High-Throughput Multi-dimensional Scaling (HiT-MDS) for cDNA-Array expression data // 15th Conference (International) on Artificial Neural Networks: Biological Inspirations Proceedings, 2005. Vol. 1. P. 625–633. doi: http://dx.doi.org/10.1007/11550822_97

Поступила в редакцию 20.12.15