Identification of Integrated Rating Mechanisms on Complete Data Sets

Nikolay Korgin, Vladimir Sergeev

Identification of Integrated Rating Mechanisms on Complete Data Sets

Nikolay Korgin[1[0000-0003-1744-3011]] and Vladimir Sergeev[1[0000-0002-7948-8656]]

[1] V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, Russia
nkrogin@ipu.ru

Abstract. An approach to the selection of structure for integrated rating mechanism to be identified given complete learning dataset is suggested. Theoretical assertions and derived from them constructive algorithm for full binary tree selection are described. Challenges for the extension of the approach suggested to incomplete data sets are outlined.

Keywords: Integrated Rating Mechanism, System Identification, Discrete data analysis.

1 Introduction

1.1 Integrated rating mechanisms

Integrated Rating Mechanisms (IRM) were introduced as multidimensional assessment and ranking systems for management and control in organizational and manufacturing systems in the Soviet Union in earlier 80th of the previous century (for example ACCORD for electronic industry, see [1]) and are implemented nowadays too ([2]-[6]). IRM usually applied for ordinal ranking (or classification) with a predetermined number of classes of a finite set of multicriteria alternatives (discrete data analysis [7]. The key components of IRM are full binary tree and convolution matrices, which allow obtaining integrated assessment (IA) based on the values of several parameters.

The typical approach to the identification of parameters of IRM in order to implement assumes iterative interaction with decision-makers (see [2],[6]). But nowadays there exist inquiry for developing of learning procedures for IRM which are common for AI algorithms identification tasks. Recently several approaches were suggested that allow to identify convolution matrices given particular binary tree – [8], [9]. While the first one allows to conduct identification task only for special full binary trees – so called sequential, the later, based on quite popular nowadays in AI algorithms – one-hot encoding (see, for example, [10]) allows one to construct algorithms for identification of IRM for any full binary tree.

Yet, due to large number of binary trees the task for predicting, what tree should be considered in process of identification is very actual problem. In this paper we suggest an approach that allows to select binary trees, for which problem of identification may be solved.

1.2 Basic notations and definitions

Let us introduce main notations on the basis of [8]. There is a finite set of indicators $L \subset N$, $|L| = l$ on the basis of their values ordinal assessment of some object or ranking of several objects should be performed. For the task of identification of IRM we will initially assume that for each indicator $i \in L$ the finite set $K_i \subset N$ of its possible values $k_i \in K_i$ is given and the tuple $k = (k_1, \dots, k_l)^T$ describes any possible state of assessed objects. There is also a finite set $K_L \subset N$ of possible integrated values (ranks or classes) $k_L \in K_L$ for any possible k.

Definition 1. IRM with a full binary tree and matrix convolutions is mapping $w(\cdot): \prod_{i\in L} K_i \to K_L$, for which indicators from L are leaves of a full binary tree – oriented graph $G = (V, E)$:

1.$V = L \cup \hat{L}$, $\hat{L} = \{l+1, \dots, 2l-1\}$,

2.$E = \{e_{ij}\} \subseteq V \times V$,

a.$\forall i \in V \backslash \{2l-1\}\ \exists!\, j \in \hat{L} \backslash \{i\}$: $e_{ij} = 1$, $\forall t \in V \backslash j\ e_{it} = 0$,

b.$\forall j \in L\ \forall i \in V\ e_{ij} = 0$,

c. $\forall j \in \hat{L} \exists!\, \{r, c\} \in V \backslash \{j\} \times V \backslash \{j\}$:$e_{rj} = 1$,$e_{cj} = 1$.

And $\forall j \in \hat{L}$ (inner node of the tree including root)

1. a finite set $K_j \subset N$ of its possible values $k_j \in K_j$, $K_{2l-1} = K_L$

2. a convolution matrix $M_j = [m_{jrc} \in K_j]_{r\in\{0,\dots,|K_l-1|\}, c\in\{0,\dots,|K_r-1|\}}$, $\{r, c\} \in V \backslash \{j\} \times V \backslash \{j\}$:$e_{lj} = 1$,$e_{rj} = 1$ are determined. □

Given some IRM $w(\cdot)$ we will denote set of all its convolution matrices as $M_w = \{M_j\}_{j\in\hat{L}}$

In this paper, we will restrict our attention to uniform-scaled IRM, such that $\forall j \in V$ $K_j = K_L$.

Given $L \subset N$ let us denote $\Gamma_2(L)$ - the set of all full binary trees with leaves from L, $IRM_{L,2}$ - the set of all IRM for with any particular full binary tree $G \in \Gamma_2(L)$, $IRM_{L,G} \subseteq IRM_{L,2}$ - the set of all IRM with such tree.

For the learning problems let us denote $q = (k, k_L)$ - a tuple of one learning example, $Q \subset K \otimes K_L$, $K = \prod_{i\in L} K_i$ - a learning set (of provided examples). The learning set is consistent if $\forall \{q, \tilde{q}\} \subseteq Q\ k \neq \tilde{k}$. The learning set is complete if $\forall k \in K\ \exists q \in Q$: $q = (k, k_L)$. The learning set is uniform-scaled if $\forall i \in L\ K_i = K_L$.

Given some arbitrary $Q \subset K \otimes K_L$, it is possible to define the following key notations concerning identification problems. The first one is the implementation problem:

Definition 2. An $w(\cdot) \in IRM_{L,2}$ implements Q, iff $\forall q \in Q\ w(k) = k_L$ □

Let us denote $IRM_{L,2}(Q)$ - set of all IRM, that implement Q, $IRM_{L,G}(Q)$ - set of all IRM, that implement Q and are based on some full binary tree $G \in \Gamma_2(L)$. Then if $IRM_{L,2}(Q) \neq \emptyset$ then Q is IRM-implementable, if $IRM_{L,G}(Q) \neq \emptyset$ then Q is IRM-implementable with structure G.

Also, definition 2 may be narrowed up to one particular learning example - $w(\cdot) \in IRM_{L,2}$ implements some $q \in Q$ iff $w(k) = k_M$ and in the same manner $IRM_{L,2}(q)$ and $IRM_{L,G}(q)$ are defined.

In case if $IRM_{L,2}(Q) = \emptyset$ or $IRM_{L,G}(Q) = \emptyset$ it is possible to state the approximation problem. Given particular $w(\cdot) \in IRM_{L,2}$ let us denote $Q_w = \{q \subseteq Q : w\,(k) = k_L\}$.

For any arbitrary $w(\cdot) \in IRM_{L,2}$ let us denote $U_Q(w) = \#Q_w/\#Q$ to be quality of approximation with maximum value 1 if $w(\cdot) \in IRM_{L,2}(Q)$. Then the approximation problem is to find $w^*(\cdot) \in \underset{w \in IRM_{M,2}}{Argmax} U_Q(w)$. And in the case, if Q is IRM-implementable, then any correct solution of the approximation problem should yield $w(\cdot) \in IRM_{L,2}(Q)$.

Given some $G \in \Gamma_2(L)$ the same problem may be stated in the restricted way – to find $w^*(\cdot) \in \underset{w \in IPM_{M,G}}{Argmax} U_Q(w)$.

Problems stated above have combinatorial nature and it may be very difficult to solve them. The $|\Gamma_2(L)| = (2l-3)!!$ [11]. That is why possibility to predict, what G should be selected or should not be considered is quite important.

2 Sensitivity and implementation via IRM

2.1 Equivalency sets

Now let us introduce an idea of sensitivity of learning set Q or function described by this set f_Q to some subset of it's indicators.

Given some arbitrary subset of indicators $\tilde{L} \subseteq L$ let us denote $k = (k_{(\tilde{L})}, k_{(L \setminus \tilde{L})})$ - splitting of indicators tuple k for some arbitrary learning example $q = (k, k_L)$ into two tuples and divide the whole set of possible values of indicators from this set $K_{(\tilde{L})} = \prod_{i \in \tilde{L}} K_i$ into a number of nonintersecting subsets $K_{\tilde{L}}(Q) = \{\kappa \subseteq K_{(\tilde{L})}\}$ such that $\forall \{q, \tilde{q}\} \subseteq Q$ if $k_{(\tilde{L})} \in \kappa$ and $\tilde{k}_{(\tilde{L})} \in \kappa$ then $k_L = \tilde{k}_L$ $(f_Q(k) = f_Q(\tilde{k}))$ and for any $\{\kappa, \tilde{\kappa}\} \subseteq K_{\tilde{L}}$ $\exists \{q, \tilde{q}\} \subseteq Q$ such that $k_{(\tilde{L})} \in \kappa$, $\tilde{k}_{(\tilde{L})} \in \tilde{\kappa}$ and $k_L \neq \tilde{k}_L$ $(f_Q(k) \neq f_Q(\tilde{k}))$. We will denote each such set κ to be an equivalency set.

Let us denote such construction $K_{\tilde{L}}(Q)$ – composition of equivalency sets for subset of indicators $\tilde{L}$

It is quite obvious, that $|K_{\tilde{L}}(Q)| \in \{1, \ldots, \prod_{i \in \tilde{L}} |K_i|\}$. Also it is quite clear, that if for some Q, and $\tilde{L} \subseteq L$ $|K_{\tilde{L}}(Q)| = 1$, then any variable from this subset is insignificant to Q (or f_Q). And the higher $|K_{\tilde{L}}(Q)|$ - the more sensitive Q to this subset of indicators. It also worth to mention that $|K_L(Q)| = K_L$ while for some $\tilde{L} \subseteq L$ it may be that $|K_{\tilde{L}}(Q)| > K_L$.

Let's consider some examples of counting of equivalence groups. An example of learning set Q on three indicators:

Table 1. Examples on full learning set on three indicators

# of example	Indicator			Assessment
	l1	l2	l3	
1	0	0	0	0
2	1	0	0	1
3	0	1	0	1

4	1	1	0	0
5	0	0	1	0
6	1	0	1	0
7	0	1	1	0
8	1	1	1	0

For this example, we have the following equivalency sets:

Table 2. Equivalency sets for full learning set on three indicators

$\widetilde{Li}$	l1	l2	l3	l1l2	l1l3	l2l3
$\lvert K_{\tilde{L}i}(Q)\rvert$	2	2	2	2	3	3

Let’s consider another example, learning set Q with four indicators:

Table 3. Examples on full learning set on four indicators

# of example	Indicator				Assessment
	l1	l2	l3	l4	
1	0	0	0	0	0
2	1	0	0	0	1
3	0	1	0	0	1
4	1	1	0	0	0
5	0	0	1	0	0
6	1	0	1	0	0
7	0	1	1	0	0
8	1	1	1	0	0
9	0	0	0	1	0
10	1	0	0	1	0
11	0	1	0	1	0
12	1	1	0	1	0
13	0	0	1	1	0
14	1	0	1	1	0
15	0	1	1	1	0
16	1	1	1	1	0

For this example, we have the following equivalency sets:

Table 4. Equivalency sets for full learning set on four indicators

$\widetilde{Li}$	l1	l2	l3	l4	l1l2	l1l3	l1l4

$\lvert K_{\widetilde{Li}}(Q)\rvert$	2	2	2	2	2	3	3
$\widetilde{Li}$	l2l3	l2l4	l3l4	l1l2l3	l1l2l4	l1l3l4	l2l3l4
$\lvert K_{\widetilde{Li}}(Q)\rvert$	3	3	2	2	2	3	3

As you can see in table above it may be that $|K_{l1}(Q)| = 2$ and $|K_{l3}(Q)| = 2$, but $|K_{l1l3}(Q)| = 3$. Also $|K_{l1l2}(Q)| = 2$, $|K_{l3l4}(Q)| = 2$ and $|K_{l2}(Q)| = 2$, $|K_{l4}(Q)| = 2$, but $|K_{l2l3l4}(Q)| = 3$, when $|K_{l1l2l3}(Q)| = 2$, $|K_{l1l2l4}(Q)| = 2$. On the figure below, we illustrate the resulting groups by implemented full binary trees on set Q from table 3.

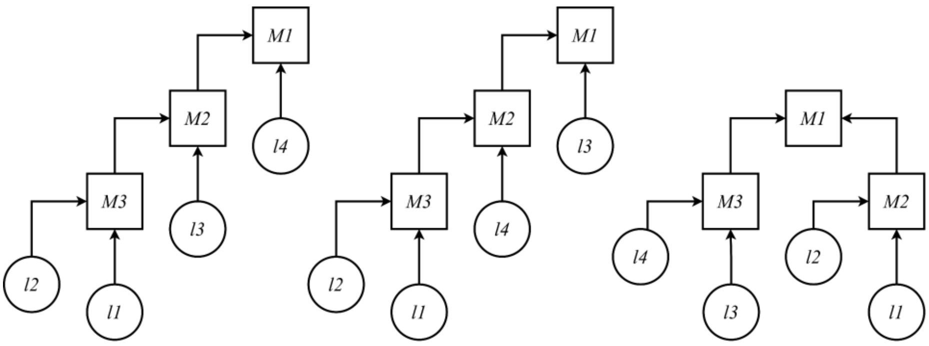

Fig. 1. Full binary trees implementing described learning set Q with four indicators

Now let us apply notions introduced above to IRM implementation problem.

For any arbitrary $G \in \Gamma_2(L)$ let us denote its indicators decomposition structure $\Lambda(G) = \{L_i\}_{i\in\{1,..,l-1\}}$ such that $\forall i \in \{1,\dots,l-1\}$ $L_i \subseteq L$ is a set of leaves (indicators) of subtree with root in node i. That is $L_1 = L$, and $\forall i \in \{1,\dots,l-1\}$ $|L_i| \geq 2$.

It is obvious that there is no explicit monotonicity $K_{\tilde{L}}$ of groups of indicators by nesting. But we can say for sure about equivalency that if $\tilde{L} \subseteq L$ $|K_{\tilde{L}}| = 1$, then $\forall \hat{L} \subseteq \widetilde{L}$ $|K_{\hat{L}}| = 1$. In addition, the following statements are true.

Assertion 1. Given some arbitrary complete Q and $G \in \Gamma_2(L)$. Then for any some $L_i \in \Lambda(G)$ such that $L_i \geq 2$ and its subgroups $\{L_{ir}; L_{ic}\} \subset \Lambda(G)$: $L_{ir} \cup L_{ic} = L_i$. For any admissible values of indicators from L_{ir} - $k_{(L_{ir})}$, $\tilde{k}_{(L_{ir})}$and from L_{ic} - $k_{(L_{ic})}$, $\tilde{k}_{(L_{ic})}$such that $\{k_{(L_{ir})}, \tilde{k}_{(L_{ir})}\} \in \kappa$, where $\kappa \in K_{L_{ir}}(Q)$, and $\{k_{(L_{ic})}, \tilde{k}_{(L_{ic})}\} \in \hat{\kappa}$, where $\hat{\kappa} \in K_{L_{ic}}(Q)$ it follows that values of corresponding indicators from L_i $k_{(L_i)} = (k_{(L_{ic})}, k_{(L_{ir})})$ and $\tilde{k}_{(L_i)} = (\tilde{k}_{(L_{ic})}, \tilde{k}_{(L_{ir})})$ belong to same indifferences set $\{k_{(L_i)}, \tilde{k}_{(L_i)}\} \in \kappa'$.

Assertion 2. In case of uniform-scale implementation for complete learning set Q $IRM_{L,G}(Q) \neq \emptyset$ iff $\forall i \in \{1,\dots,l-1\}$ $\left|K_{L_i}(Q)\right| \leq |K_L|$.

Assertion 2 allows to construct algorithm for selecting $G \in \Gamma_2(L)$ such that $IRM_{L,G}(Q) \neq \emptyset$ in case if Q is implementable.

2.2 Algorithm of tree selection in case of complete learning set

First of all, obtaining equivalence groups allows us to remove all structures $G \in \Gamma_2(L)$ that include combinations with the number of equivalence groups exceeding the scale, $|K_{\tilde{L}i}(Q)| \geq K_L$. This allows us to dramatically decrease number of structures to be considered with any algorithm of IRM identification for particular $G \in \Gamma_2(L)$ [8,9]. The performance of equivalence testing of one subset of indicators $\tilde{L}$ is proportional $(K_L)^l$. Second, based on the equivalence groups within the scale $|K_{\tilde{L}i}(Q)| \leq K_L$, we can use the algorithm to select obviously allowed structures.

Algorithm: If, as in a result of the analysis of equivalence groups, we have allowed combinations of leaves, then we begin our consideration with groups consisting of three leaves $|L_i| = 3$. For such groups, we check the possibility of their decomposition into groups of allowed pairs of leaves and individual leaves. Using only permitted combinations of leaves reduces the number of structures considered. Thus, we compile a list of checked groups of three leaves. Next, go to the groups of four leaves and repeat the decomposition test based on the groups selected above. Thus, we check for the possibility of decomposition of all groups with $|L_i| \geq 2$ to the maximum size $|L_i| = l$ leaves in the group. If it is not possible, then such groups are excluded from list. These iterations should be repeated until full set of leaves may be decomposed in number of subgroups remained in the list of allowed groups. Then any structure $G \in \Gamma_2(L)$ such that its $\Lambda(G)$ is included in list of allowed leaves groups should be considered in IRM's identification task for this learning set Q on the basis of approaches, suggested in [8].

2.3 Challenges for implementation of incomplete learning sets

While the approach suggested in previous section works perfect with complete learning sets, for practical application it is important to develop methods for incomplete learning sets and there are still number of challenges to deal with. In particular, Assertion 1 is not correct in case of incomplete sets. Which leads to challenges at stage of assessment on number of equivalence groups given particular set of indicators and at the stage of application of the algorithm, described above. This is an actual problem to be solved in the next stage of development of the approach described above.

Author 1 acknowledges for partial funding RFBR grant 19-29-07525 (paragraphs 1, 2.1). Author 2 acknowledges for partial funding RSF grant 17-78-20047 (paragraphs 2.2-2.3).

References

1. Gorelikov, N.: Designing a sectoral model for management of development and manufacture of new products. Avtomatika i Telemekhanika (5), 63-70 (1984).
2. Burkov V., Novikov D., Shchepkin A.: Control Mechanisms for Ecological-Economic Systems. Springer, Berlin (2015).

3. Korgin N., Rozhdestvenskaya S.: Concordant Approach for R&D Projects' Evaluation and Ranking for Formation of Programs. In IEEE 11th International Conference on Application of Information and Communication Technologies (AICT), pp. 1-5. IEEE, (2017).
4. Shchepkin A.: Application of Integrated Mechanism in Financing Project Works. In 13th International Conference" Management of large-scale system development"(MLSD), pp. 1-4. IEEE, (2020).
5. Zheglova Y., Titarenko B.: Methodology for the integrated assessment of design solutions for foundation pit fences based on the theory of active systems. In IOP Conference Series: Materials Science and Engineering, vol. 869, p. 052012. IOP Publishing, (2020).
6. Burkov V. et al.: Models and Management Structure for the Development and Implementation of Innovative Technologies in Railway Transportation. I. Mechanisms of Priority Projects Selection and Resource Allocation. Automation and Remote Control (81), 1316-1329 (2020).
7. Mariel P. et al.: Environmental valuation with discrete choice experiments: Guidance on design, implementation and data analysis. Springer Nature, Berlin (2021).
8. Burkov V., Korgin N., Sergeev V.: Identification of Integrated Rating Mechanisms as Optimization Problem. In 13th International Conference" Management of large-scale system development"(MLSD), pp. 1-5. IEEE, (2020).
9. Alekseev A.: Identification of Integrated Rating Mechanisms Based on Training Set. In 2nd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA), pp. 398-403. IEEE, (2020).
10. Yu L. et al.: Missing Data Preprocessing in Credit Classification: One-Hot Encoding or Imputation? Emerging Markets Finance and Trade, 1-11 (2020).
11. Barnett J. et al.: Darwin Meets Graph Theory on a Strange Planet: Counting Full n-ary Trees with Labeled Leafs. Alabama Journal of Mathematics (35), 16-23 (2010).