ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20190615044223
Original scientific paper

Index System Reduction Method Based on the Index Similarity

Ai WANG, Xuedong GAO

Abstract: Multi-attribute decision making (MADM) always suffers from the result inconsistency and computational complexity problem, due to numbers of redundant and relational attributes (indexes) of the initial evaluation index system. Therefore, this paper studies the index system (IS) reduction problem through selecting the most representative indicator from each index subsystem after the IS structure partition. First, we propose and demonstrate the Index Subsystem Judgement theorem to improve the efficiency of the classic system structure partition algorithm. Second, an algorithm of index system reduction based on the index similarity (ISRS) is put forward. The ISRS is able to reduce the index quantity while still keeping the index meaning. Third, we define the direction loss rate to measure the evaluation ability loss of the IS during reduction. The algorithm is tested for a synthetic dataset to compare the proposed ISRS with different index reduction algorithms, followed by an extensive experimentation with a real-world financial dataset. Experiment results illustrate that our proposed method is able to obtain more accessible and available reduction results in practice.

Keywords: direction loss rate; Index Subsystem Judgment theorem; index system reduction; index similarity

1 INTRODUCTION

Multi-attribute decision making (MADM) is a sub-field of operations research, concerned with selecting the best alternative through the evaluation of the whole set of attributes which are hard to quantify, incommensurable or incomparable [2, 35]. A number of redundant or relational attributes (indexes) might increase the potential internal inconsistency and computational complexity of the MADM methods, such as analytic hierarchy process (AHP) [37, 39]. To deal with this drawback, an appropriate index reduction should be implemented.

Since an index system is exactly a system which consists of different indexes (elements) with specific structure (relation), the index system reduction problem could be definitely transformed to the system structure partition problem, that is selecting the most representative index from each index subsystem.

Formally, let S denote the initial index system, the task of index system reduction is to partition S into several index subsystems S_i and utilize one index X_{ik} to replace each subsystem, where (1) $S_i \neq \emptyset$, (2) $\cup S_i = S$, (3) $S_i \cap S_j = \emptyset$. Here,$X_{ik} \in S_i \subseteq S$ is an index and the final reduced index system $O \subseteq S$.

Tab.1 is a common evaluation index system on students' learning performance, including every course score as well as the total and average grade. It can be seen that not only there is the direct linear relation between the total and average grade (that is if one of the Total and Average is known, the other could be calculated through simply multiplying a coefficient), also the performance of different students on the same course shares some similarity, such as Math and Physics.

Obviously, in order to improve the evaluation efficiency of the index system in Tab.1, it is better to just remove one index from each subsystem (i.e., {Math, Physics} and {Total, Average}), which successfully reduces the index quantity while still keeping the index meaning. We study the feature of these two subsystems and find out that indexes in the same subsystem are much more similar to each other than to those in other subsystems, which illustrates that the partition principle of index subsystems is on the basis of the index similarity.

Therefore, this paper focuses on the index system reduction problem based on the index similarity. The main contributions are as follows. First, we propose and demonstrate the Index Subsystem Judgement theorem and its inference, which can improve the efficiency of the system structure partition process, even be valid for other problems (e.g. high dimensional data pre-processing, knowledge discovery). Second, an algorithm of index system reduction based on the index similarity (ISRS) is also proposed. The ISRS is able to reduce the index quantity while still keeping the index meaning compared to traditional index reduction algorithms, which shows great advantages in practice. Third, we define the direction loss rate to measure the evaluation ability loss of the index system, which could verify and evaluate the results obtained by the ISRS.

The rest of the paper is organized as follows. Section 2 presents the previous works related to this research. Section 3 presents the methodology of the index system reduction, including the theorem proof and algorithm. Section 4.1 conducts several comparison experiments related to different index system reduction algorithms and similarity evaluation indices on a synthetic dataset. Besides, experiments in Section 4.2 further verify the effectiveness and stability of the proposed method on a real financial data set. The paper is concluded in Section 5.

Table 1 An index system on students' learning performance

Student	English	Chinese	Math	Physics	Total	Average
Tom	98	90	92	91	371	92.75
Sherry	96	74	87	88	345	86.25
Bill	57	87	45	49	238	59.50
Jack	84	79	60	60	283	70.75
Mary	47	92	61	59	259	64.75

2 LITERATURE REVIEW

2.1 Index System Reduction Algorithms

This section reviews the conventional index system reduction algorithms, such as the principal component analysis (PCA) [4], rough set (RS) theory [12] and independent component analysis (ICA) [6].

Tab. 2 describes the comparison of the PCA, RS theory and the proposed algorithm ISRS (see Section 3). Although all three algorithms are able to solve the reduction problem, the emphasis is somewhat different [8, 15]. For example, the RS theory mainly focuses on the attribute reduction problem [23, 28] and shows much potential in the multi-label learning [29, 36], while the PCA is usually applied to the dimension reduction problem in the machine learning [9, 16] and multi-objective optimization (MOO) [18, 20]. What's more, the RS theory reduces attributes based on the partition of equivalence relation, which can be accomplished through the evaluation metric [30, 31], like mutual information and information entropy [10, 38]. In general, the RS theory achieves better performance on the categorical attribute dataset [1, 11]. The PCA reduces dimensions based on the variance additivity of irrelevant principal components [21, 22], which can be evaluated by the contribution rate [19, 27]. The purpose of the ICA is similar to the PCA, which aims to obtain independent components for dimension reduction [25, 26]. However, the proposed ISRS in this paper achieves the index reduction based on the index similarity, especially for the multi-attribute decision making (MADM). Also, direction loss rate is put forward, in order to measure the evaluation ability loss of reduced index system.

Besides, the results of three algorithms are quite different, that is the decision table (of RS theory), new principal components (of PCA) and index system (of ISRS) [3, 33, 34]. From the perspective of ISRS, both independent components (of ICA) and principal components are the combination of initial indexes, which means compared to the ISRS, the PCA and ICA have the same effect. Since this paper aims to study the index system reduction problem of the numerical dataset, the comparison experiments of different algorithms in Section 4.1 are mainly conducted between the PCA and ISRS.

Table 2 The comparison of index system reduction algorithms

Method	RS theory	PCA	Proposed ISRS
Purpose	Attribute reduction	Dimension reduction	Index reduction
Principle	Equivalence relation	Variance additivity	Index similarity
Evaluation Metric	Mutual information	Contribution rate	Direction loss rate
Result	Decision table	Principal components	Index system
Typical Field	Multi-label learning	Multi-objective optimization (MOO)	Multi-attribute decision making(MADM)

2.2 Similarity Evaluation Indices

Similarity evaluation indices is a popular research field of data mining, and are wildly utilized in clustering and classification algorithms [5]. We review three typical metrics as follows.

Give an index system $S=\{X_1, X_2, \dots, X_n\}$, let $\overrightarrow{x_k}=(v_{k1}, v_{k2}, \dots, v_{km})$ represents the sample vector of index X_k, and $\overrightarrow{x_l}=(v_{l1}, v_{l2}, \dots, v_{lm})$ represents the sample vector of index X_l. The Cosine distance d, Euclidean distance $\dot{d}$, Mahalanobis distance $\ddot{d}$ of index X_k and X_l is:

$$d(X_k, X_l)=1-\overrightarrow{x_k}\cdot\overrightarrow{x_l}/(|\overrightarrow{x_k}||\overrightarrow{x_l}|) \quad (1)$$

$$\dot{d}(X_k, X_l)=\sqrt{\sum_{t=1}^{m}(v_{kt}-v_{lt})^2} \quad (2)$$

$$\ddot{d}(X_k, X_l)=\sqrt{(\overrightarrow{x_k}-\overrightarrow{x_l})^T\Sigma^{-1}(\overrightarrow{x_k}-\overrightarrow{x_l})} \quad (3)$$

where Σ is the covariance matrix of index system S.

Cosine distance measures the direction difference of index vectors, Euclidean distance measures the length difference of index vectors, and Mahalanobis distance measures the covariance distance of index vectors. Although Mahalanobis distance overcomes the correlation between indicators and is scale-invariant, it strictly requires the number of indexes must be larger than the number of samples (because we want to calculate the similarity of different indexes not samples), which is not consistent with the index dataset in practice. Thus, the comparison experiments of different indices in Section 4.1 are mainly conducted between the Cosine distance and Euclidean distance.

3 INDEX SYSTEM REDUCTION METHOD BASED ON THE DIRECTION LOSS RATE

This section studies the theoretical framework and algorithm of the index system reduction problem. Since the index system reduction is exactly the system structure partition in essence, the classic method of the system structure partition, that is the interpretative structural modeling (ISM), is applied to our research [14, 32].

Fig. 1 shows the principle of the index system reduction method based on the index similarity. There are four major phases. First, calculate the index similarity through sufficient samples, in order to identify the relation between different indicators. Then, divide the initial index system into several index subsystems by means of the improved ISM. After that, replace each subsystem with its most representative index, and obtain the reduced index system. Finally, calculate the evaluation ability loss during reduction via the proposed metric, direction loss rate.

According to the example of students' learning performance in Section 1, since the indexes with linear relation ought to be reduced (like Total and Average), the direction difference of different indexes plays a more significant role than the length difference. Thus, we take the Cosine distance as the similarity evaluation indices (see Eq. (1)).

Refer to the ISM, the final similarity of different elements depends on the overall relation, not only limited to the direct relation (i.e., Cosine distance) [7]. Consequently, we judge the relationship between indexes through the overall relation in the reachable matrix. Given index X_k and X_l, the judgement function of the overall cosine distance $\bar{d}(X_k, X_l)$ is:

$$J(X_k, X_l)=\begin{cases}1, & \bar{d}(X_k, X_l)\le\delta\\ 0, & \bar{d}(X_k, X_l)>\delta\end{cases} \quad (4)$$

Where δ represents the similarity threshold, and $J(X_k, X_l)=1$ means X_k is definitely similar to X_l, that is X_k can reach X_l; otherwise, $J(X_k, X_l)=0$ means X_k is not similar to X_l, that is X_k cannot reach X_l.

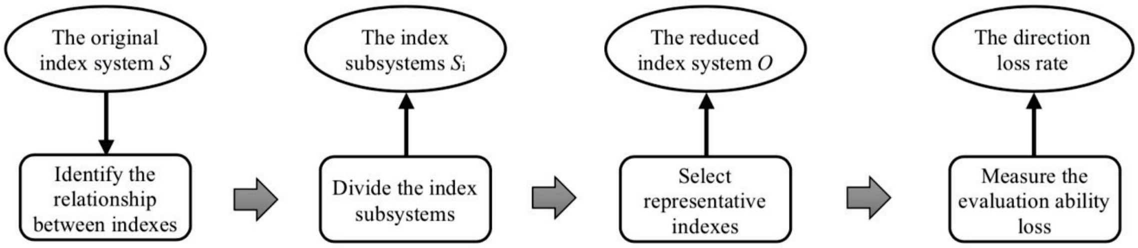

Figure 1 The principle of the index system reduction method based on the index similarity

Theorem 1 (Index Subsystem Judgment theorem). In a reachable matrix, all the indexes with the same reachable set or the same antecedent set only belong to one index subsystem.

Proof. Given an index system S and index $X_k, X_l \in S$.

Assume that X_k, X_l have the same reachable set, that is $A(X_k) = A(X_l)$.

Because $X_k \in A(X_k)$ and $A(X_k) = A(X_l)$, $X_k \in A(X_l)$. Consequently, the index X_l can reach X_k. In the same way, the index X_k can reach X_l.

That X_k, X_l reach each other means they belong to one strong connected domain (subsystem).

Hence, all the indexes with the same reachable set belong to one subsystem.

If index X_l also belongs to another subsystem that contains the index X_p, while X_k does not, then is $X_p \in A(X_l)$ and $X_p \notin A(X_k)$.

However, $X_p \in A(X_l)$ and $X_p \notin A(X_k)$ contradicts the assumption $A(X_k) = A(X_l)$.

Thus, all the indexes with the same reachable set only belong to one subsystem.

The antecedent set can be used to prove the same conclusion.

Algorithm 1: Subsystem partition algorithm based on the merge of elements (SP).

Input: The adjacent matrix A of a system $S = \{X_k | k \in [1, n]\}$ and partition parameter δ.
Output: The subsystems $S_i (S = \cup S_i)$.

```
R=A. Overall Relation(δ) // see Eq.4
for 1 ≤ k ≤ n do
for 1 ≤ l ≤ n do
if R_kl = 1 then
A(X_k) = A(X_k) ∪ {X_l}
A(X_l) = A(X_l) ∪ {X_k}
end if
    end for
  end for
for S ≠ ∅ do
for all X_k ∈ S do
S_i = {X_k}
if A(X_k) = A(X_l) then
S_k = S_k ∪ {X_l}
end if
return S_i
S.delete(S_i)
 end for
 end for
```

Inference. In a reachable matrix, the indexes with different reachable sets are bound to have different antecedent sets, vice versa.

We improve the ISM via the Theorem 1, and the pseudo code is shown in Algorithm 1. The time complexity of the SP is $O(n^2/2)$, where n is the total number of indexes, which obtains an obvious improvement towards the classic algorithm $O(n^3)$.

Phase 3 (that is select a representative index) is the core index reduction process after the preparation Phase 1 and 2, which share the same essence to the attribute selection problem in data mining research field [5]. Hence, we take the wildly-used attribute selection metric, information entropy, as the measurement of index evaluation ability during reduction (see Eq.5) [24].

Given the index vector $\overrightarrow{x_k} = (v_{kt})$, $(t \in [1, m])$ of index X_k, v_{kt} represents the sample value (of sample O_t on index X_k), and $p(v_{kt})$ represents the probability of value v_{kt}. The evaluation ability of index X_k is:

$$H(X_k) = -\sum_{t=1}^{m} p(v_{kt}) \log(p(v_{kt})) \tag{5}$$

where $0 < p(v_{kt}) \le 1$, and $\sum_{t=1}^{m} p(v_{kt}) = 1$.

As for the entropy decrease effect, the index with larger information entropy should be abandoned due to the poor information value; while the index with smaller information entropy should be reserved due to the rich information value [17].

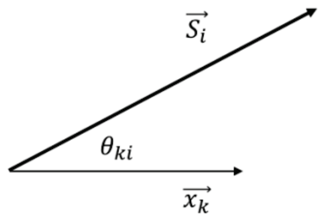

Figure 2 The index replacement process during reduction

Moreover, it is also necessary to measure the evaluation ability loss of the index system after reduction. Fig.2 shows the index replacement process on one index subsystem. Let $\overrightarrow{x_k}$ represent the vector of index X_k in the index subsystem S_i, and $\overrightarrow{S_i}$ represent the resultant vector of subsystem S_i, that is $\overrightarrow{S_i}$ consists of all the index vectors in S_i. If using index X_k to replace the whole subsystem S_i, the evaluation direction of the index subsystem changes from $\overrightarrow{S_i}$ to $\overrightarrow{x_k}$. Thus, the evaluation ability loss of S_i during reduction is exactly the evaluation direction θ_{ki}.

Definition 1 (Subsystem Direction Loss Rate). Let $S_i = \{X_k | k \in [1, u]\}$ represent an index subsystem, $\overrightarrow{x_k} = (v_{k1}, v_{k2}, \ldots, v_{km})$ represent the vector of index X_k in S_i, and $\overrightarrow{S_i}$ represent the resultant vector of subsystem S_i. If using index X_k to replace subsystem S_i, the direction loss rate λ_i of index subsystem S_i is:

$$\lambda_i = \theta_{ki}/\pi \tag{6}$$

$$cos\theta_{ki} = \frac{\overrightarrow{x_k}\cdot\overrightarrow{S_i}}{|\overrightarrow{x_k}||\overrightarrow{S_i}|} = \frac{\sum_{t=1}^{m}(v_{kt}\sum_{j=1}^{u}v_{jt})}{\sqrt{\sum_{t=1}^{m}(v_{kt})^2}\sqrt{\sum_{t=1}^{m}(\sum_{j=1}^{u}v_{jt})^2}} \tag{7}$$

where $\theta_{ki} \in [0,\pi]$ is the direction loss degree of subsystem S_i when replaced by index X_k.

Algorithm 2: Index System Reduction algorithm based on the index similarity (ISRS).

Input: An index system $S = \{X_k | k \in [1,n]\}$ and similarity threshold δ.
Output: The reduced index system O, and its average direction loss rate $\bar{\lambda}$.

A=S. *indexSimilarity*// see Eq.1
S_i = SP(A,δ)// see Algorithm 1, and $S = \bigcup_{i=1}^{r} S_i$
for $1 \le i \le r$ **do**
if $|S_i| > 1$ **then**
for all $X_l \in S_i$ **do**
H(X_k) = min(H(X_l))
end for
$S_i = \{X_k\}$
λ_i = arccos$(cos\theta_{ki})/\pi$// see Eq.6,7
else
$\lambda_i = 0$
end if
end for
$O = \bigcup_{i=1}^{r} S_i$
$\bar{\lambda} = \sum_{i=1}^{r} \lambda_i |S_i|/m$// see Eq.8
return $O, \bar{\lambda}$

Since there is no similarity between different index subsystems, the evaluation ability loss of the whole index system can be measured through calculating the weighted average of the direction loss rate from each subsystem.

Definition 2 (System Average Direction Loss Rate). Let $S_i (i \in [1,r])$ represent the subsystems of index system S, The average direction loss rate $\bar{\lambda}$of index system S is:

$$\bar{\lambda} = \sum_{i=1}^{r} \lambda_i |S_i|/|S| \tag{8}$$

where $|S_i|$ represents the index quantity of subsystem S_i, and $|S|$ represents the index number of system S.

Finally, we propose the index system reduction algorithm based on the index similarity (ISRS), and the pseudo code is shown in Algorithm 2. The time complexity of the ISRS is $O((n^2+nm)/2)$, where n is the total number of indexes and m is the number of samples.

4 EXPERIMENT AND RESULTS ANALYSIS

4.1 Comparison Experiment of Different Algorithms

In order to verify the effectiveness of the proposed ISRS, we used random number generation to create the dataset with eight indexes and fourteen samples (see Tab.3). Since we limited the value range, all the indexes are under the same scale. Therefore, no further data normalization should be taken.

Comparison experiments are implemented in this section, including different index system reduction methods, i.e., ISRS and PCA (see Section 2.1), as well as different similarity evaluation indices, i.e., Cosine distance and Euclidean distance (see Section 2.2). All the experiments were coded in Matlab 7.8 and run on a personal computer with Windows 7.

Table 3 An index system S with 8 indexes and 14 samples

O	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
O_1	40.4	37.05	7.2	6.1	8.3	8.7	2.442	20
O_2	25	19.05	11.2	11	12.9	20.2	3.542	9.1
O_3	13.2	4.95	3.9	4.3	4.4	5.5	0.578	3.6
O_4	22.3	10.05	5.6	3.7	6	7.4	0.716	7.3
O_5	34.3	17.7	7.1	7.1	8	8.9	1.726	27.5
O_6	35.6	18.75	16.4	16.7	22.8	29.3	3.017	26.6
O_7	22	11.7	9.9	10.2	12.6	17.6	0.847	10.6
O_8	48.4	20.1	10.9	9.9	10.9	13.9	1.772	17.8
O_9	40.6	25.65	19.8	19	29.7	39.6	2.449	35.8
O_{10}	24.8	12	9.8	8.9	11.9	16.2	0.789	13.7
O_{11}	12.5	14.55	4.2	4.2	4.6	6.5	0.874	3.9
O_{12}	1.8	0.9	0.7	0.7	0.8	1.1	0.056	1
O_{13}	32.3	20.85	9.4	8.3	9.8	13.3	2.126	17.1
O_{14}	38.5	13.65	11.3	9.5	12.23	16.4	1.327	11.6

Table 4 The sample value of the new reduced index system O^{PCA}

O	$\mathbf{Z_1}$	$\mathbf{Z_2}$	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8
O_1	12.0	-20.8	9.0	-3.7	-0.6	0.3	0.3	0.0
O_2	0.7	5.2	1.9	-7.9	1.5	-0.9	0.1	-0.2
O_3	-25.0	3.2	-1.7	1.0	0.2	0.5	0.3	-0.2
O_4	-15.0	-1.9	-2.7	0.4	-1.3	-0.6	0.4	0.2
O_5	5.7	-9.1	1.4	11.6	0.9	-0.3	-0.2	-0.1
O_6	23.8	10.5	0.0	1.5	1.0	0.7	0.7	0.3
O_7	-4.5	7.9	-0.9	-2.0	0.1	0.7	-0.5	-0.1
O_8	14.5	-12.3	-9.2	-0.8	0.2	0.2	-0.1	-0.3
O_9	40.7	14.4	4.3	1.8	-1.3	-0.3	-0.1	-0.2
O_{10}	-2.5	4.9	-1.4	0.8	-0.4	0.1	-0.5	0.3
O_{11}	-21.6	0.1	5.6	-3.3	0.0	0.4	-0.3	0.1
O_{12}	-38.4	5.8	2.1	3.8	-0.4	-0.3	0.2	-0.1
O_{13}	4.1	-5.5	1.3	0.0	0.6	-0.4	-0.3	0.4
O_{14}	5.5	-2.4	-9.6	-3.1	-0.6	-0.1	0.1	0.1

Table 5 Reduction process of ISRS on Cosine and Euclidean distance

Indices	Index Subsystems	O^{ISRS}	$\bar{\lambda}$
Cosine Distance	$\{X_1,X_2\}$ $\{X_3,X_4,X_5,X_6\}$ $\{X_7\}$ $\{X_8\}$	$\{X_1\}$ $\{X_5\}$ $\{X_7\}$ $\{X_8\}$	0.0183
Euclidean Distance	$\{X_1\}$ $\{X_2\}$ $\{X_3,X_4,X_5,X_6,X_8\}$ $\{X_7\}$	$\{X_1\}$ $\{X_2\}$ $\{X_8\}$ $\{X_7\}$	0.0193

Results are shown in Tab. 4 and 5. The comparison experiment results on different index reduction algorithms illustrate that (1) the reduction rate of the PCA (6/8 = 0.75) is larger than the ISRS (4/8 = 0.5). Hence, the index reduction effect of PCA is stronger than the ISRS; (2) the index meaning of the reduced index system $O^{PCA}$is much more complicated than O^{ISRS}, besides, O^{PCA} contains negative value while both the initial index system S and O^{ISRS} do not exist. Therefore, the results of ISRS is more accessible and available in practice than the PCA.

The comparison experiment results on different similarity evaluation indices illustrate that (1) the direction loss rate of Cosine distance ($\overline{\lambda_{cos}}$=0.0183) is smaller than Euclidean distance ($\overline{\lambda_{euc}}$=0.0193). Hence, as for the ISRS, Cosine distance is more accurate and effective than Euclidean distance.

4.2 Index System Reduction of Real Financial Dataset

After demonstrating the accuracy, this section further verified the stability and efficiency of the ISRS on a real financial dataset with three-hundred enterprise samples. The data structure of this financial dataset in 2015 Resset database is shown in Tab. 6. In the beginning, we conducted the data preprocessing. We identified that the No. 70-90 indexes of three-hundred enterprises are all null value attributes, which are not able to distinguish or evaluate samples. Thus, we removed these indexes from the initial index system $\tilde{S}$, and the formal index reduction experiment only focused on the No. 1-69 indicators.

Fig. 3 describes the result of index subsystems divided by the ISRS, where blue and orange represent two different subsystems respectively, and grey means every index is a subsystem. What's more, the histogram also shows the information entropy of indexes. The ISRS replaced the subsystem in blue with index No. 36, and replaced the subsystem in orange with index No. 64. Finally, we obtained the reduced financial index system $\tilde{O}$ with only forty-five indexes (see Tab. 7). It can be seen that the ISRS achieved great reduction effect on the real dataset, that the reduction rate has already reached 0.5 while keeping the average direction loss rate under 0.023.

In order to test the stability of the ISRS, we take the reduced index system $\tilde{O}$ as the standard result, and gradually cut down the data size from three-hundred to eighty. Fig. 4 shows the error rate of the ISRS under the decrease of samples. It can be seen that the tendency of this broken line presents three stages, that is the rapid descent stage (80-160), steady descent stage (160-220) and stable fluctuation stage (220-280). For the first stage, the error rate stays in a relatively high level due to too few samples involved in the ISRS calculation. For the second stage, the ISRS has already identified the potential characteristic among these enterprises, and the effect of increasing samples is not as significant as before. For the third stage, the error rate stabilizes at a very low level (0.047), which means the ISRS has converged and its index system reduction result is reliable.

Table 6 The real financial index system $\tilde{S}$ of balance sheets with 90 indexes

No.	Index	No.	Index	No.	Index
1	Monefd	31	TotNcurass	61	Surres
2	Trafinass	32	Totass	62	Retear
3	Noterecv	33	STloan	63	OrdRiskResFd
4	Accrecv	34	Trafindb	64	SHEwiomin
5	Advpay	35	Notepay	65	minSHE
6	Intrecv	36	Accpay	66	SEAdjItems
7	Othrecv	37	Advrecp	67	TotSHE
8	Invtr	38	Empsalpay	68	LEAdjItems
9	Defchr	39	Taxexppay	69	TotliaSHE
10	Ncurass1Y	40	Intpay	70	Divrecv
11	Othcurass	41	Divpay	71	Consbioass
12	CAExcItems	42	Othaccpay	72	Oilgasass
13	CAAdjItems	43	Accrexp	73	NCAExcItems
14	Totcurass	44	Ncurlia1Y	74	NCAAdjItems
15	Soldfinass	45	Othcurlia	75	AssExcItem
16	Holdinvterm	46	Totcurlia	76	AssAdjItem
17	LTrecv	47	LTloan	77	ImpLoan
18	LTequinv	48	Bdpay	78	STbdpay
19	Invrealest	49	LTpay	79	Defprcd
20	Fixass	50	Spepay	80	Specurlia
21	Constrpro	51	Estmlia	81	CLAdjItems
22	Constrmat	52	Deftaxdb	82	LExcItem
23	Dispofixass	53	OthNcurlia	83	LAdjItems
24	Prodbioass	54	SpeNcurlia	84	Treastk
25	Intanass	55	NCLAdjItems	85	Spdtrafogcursta
26	Devlpexp	56	TotNcurlia	86	Uncfinvlos
27	Goodwill	57	Totlia	87	Othres
28	LTdefchr	58	Shrcap	88	SEExcItems
29	Deftaxass	59	Capsur	89	SEOthEff
30	OthNcurass	60	SpeRes	90	LEExcItems

Table 7 The reduced financial index system $\tilde{O}$ of balance sheet with 45 indexes

No.	Index	No.	Index	No.	Index
2	Trafinass	26	Devlpexp	47	LTloan
3	Noterecv	27	Goodwill	49	LTpay
5	Advpay	28	LTdefchr	50	Spepay
6	Intrecv	29	Deftaxass	51	Estmlia
7	Othrecv	30	OthNcurass	52	Deftaxdb
8	Invtr	35	Notepay	53	OthNcurlia
10	Ncurass1Y	36	Accpay	54	SpeNcurlia
11	Othcurass	37	Advrecp	56	TotNcurlia
15	Soldfinass	38	Empsalpay	58	Shrcap
18	LTequinv	39	Taxexppay	59	Capsur
19	Invrealest	40	Intpay	60	SpeRes
21	Constrpro	41	Divpay	61	Surres
22	Constrmat	42	Othaccpay	62	Retear
24	Prodbioass	44	Ncurlia1Y	64	SHEwiomin
25	Intanass	45	Othcurlia	65	minSHE

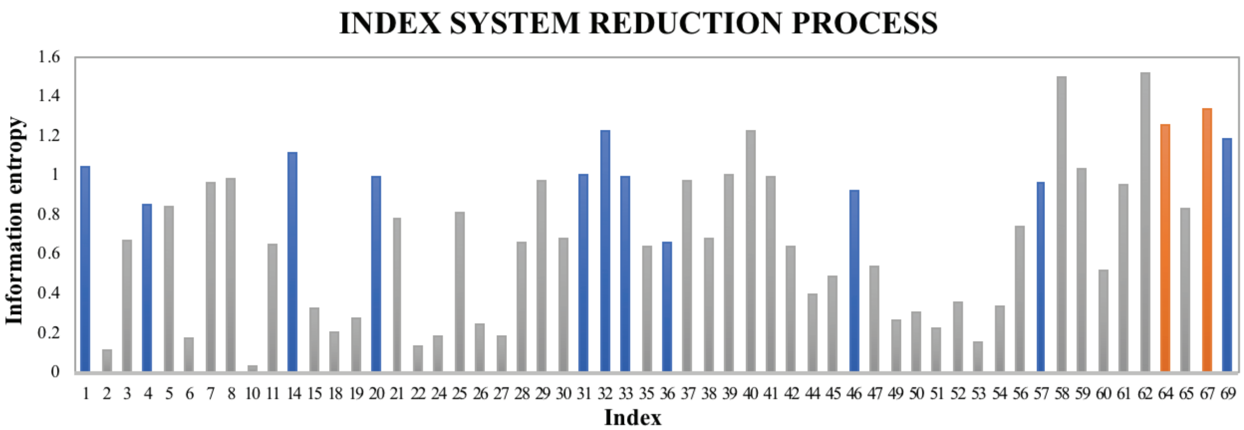

Figure 3 The information entropy and index subsystems

The experimental results on the real dataset illustrates that the ISRS is more applicable to the large-sample dataset.

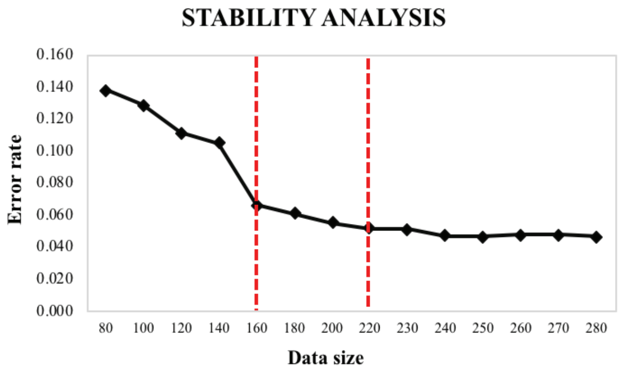

Figure 4 The error rate of the ISRS under various data size.

5 CONCLUSIONS

Index system reduction has become one of the most popular research fields since its first appearance, and is widely applied to the multi-attribute decision making (MADM). Unlike the previous approaches (like PCA) that replace indexes with new components, this paper studies the index system reduction problem from the perspective of system structure partition.

We proposed and demonstrated the Index Subsystem Judgement theorem, that improves the efficiency of the classic system structure partition algorithm, i.e., the interpretative structural modeling (ISM). An algorithm of index system reduction based on the index similarity (ISRS) was also proposed. The ISRS is able to reduce the index quantity while still keeping the index meaning, through directly selecting the most representative index from each index subsystem. Moreover, the average direction loss rate was put forward, which successfully measures the evaluation ability loss of the index system during reduction.

Experiments on both synthetic and real-world datasets illustrate that the ISRS is able to converge under the dataset with sufficient samples, and obtain more accessible and available reduction results in practice.

The future work of our study is to improve the efficiency of ISRS, especially for large-scale data analysis. Also, we will further verify the performance of our proposed method through more practical problems.

Acknowledgements

The study is supported by national natural science foundation of China (71272161) and China Scholarship Council.

6 REFERENCES

[1] Bi, Z., Xu, F., Lei, J., & Jiang, T. (2016). Attribute reduction in decision-theoretic rough set model based on minimum decision cost. *Concurrency and Computation-Practice & Experience*, *28*(15), 4125-4143. https://doi.org/10.1002/cpe.3830

[2] Castro, D. M. & Parreiras, F. S. (2018). A Review on Multi-Criteria Decision-Making for Energy Efficiency in Automotive Engineering. *Applied Computing and Informatics*. In Press. https://doi.org/10.1016/j.aci.2018.04.004

[3] Chen, J., Lin, Y., Lin, G., Li, & Y. Zhang, Y. (2017). Knowledge-Based Systems Attribute reduction of covering decision systems by hypergraph model. *Knowl. Based Syst., 118*(1), 93-104. https://doi.org/10.1016/j.knosys.2016.11.010

[4] Dubey, A. K. & Yadava, V. (2008). Multi-objective optimization of Nd: YAG laser cutting of nickel-based superalloy sheet using orthogonal array with principal component analysis. *Opt. Lasers Eng*., *46*(2), 124-132. https://doi.org/10.1016/j.optlaseng.2007.08.011

[5] Han, J. (2012). *Data Mining: Concepts and Techniques*, third ed., China Mach. Press, Beijing.

[6] Huang, J., Zhou, Z., Gao, Z., Zhang, M., & Yu, L. (2017). Aerodynamic multi-objective integrated optimization based on principal component analysis. *Chinese J. Aeronaut., 30*(4), 1336-1348. https://doi.org/10.1016/j.cja.2017.05.003

[7] Jeya, G., Sekar, V., & Vimal, K. (2016). Application of interpretative structural modelling integrated multi criteria decision-making methods for sustainable supplier selection. *J. Model. Manag., 11*(2), 358-388. https://doi.org/10.1108/JM2-02-2014-0012

[8] Kamble, P. D., Waghmare, A. C., Askhedkar, R. D., & Sahare, S. B. (2017). Multi objective optimization of turning parameters considering spindle vibration by Hybrid Taguchi Principal Component Analysis (HTPCA). *Mater. Today: Proceedings, 4*(2), 2077-2084. https://doi.org/10.1016/j.matpr.2017.02.053

[9] Kapsoulis, D., Tsiakas, K., Trompoukis, X., Asouti,V., & Giannakoglou, K. (2018). Evolutionary multi-objective optimization assisted by metamodels, kernel PCA and multi-criteria decision making techniques with application in aerodynamics. *Appl. Soft Comput., 64*(1), 1-13. https://doi.org/10.1016/j.asoc.2017.11.046

[10] Li, H., Li, D., Zhai, Y., Wang, S., & Zhang, J. (2016). A novel attribute reduction approach for multi-label data based on rough set theory. *Inf. Sci. (Ny)., 367-368*(1), 827-847. https://doi.org/10.1016/j.ins.2016.07.008

[11] Li, F., Yang, J., Jin, C., & Guo, C. (2017). A new effect-based roughness measure for attribute reduction in information system. *Inf. Sci. (Ny)., 378*(1), 348-362. https://doi.org/10.1016/j.ins.2016.08.056

[12] Lin, Y., Li, Y., Wang, C., & Chen, J. (2018). Attribute reduction for multi-label learning with fuzzy rough set. *Knowledge-Based Systems*, *152*(15), 51-61. https://doi.org/10.1016/j.knosys.2018.04.004

[13] Li, J. &Wang, X. (2016).Decision reduction for object oriented concept lattices.*Computer Engineering and Applications*. 52 (18), 154-157. https://doi.org/10.3778/j.issn.1002-8331.1512-0353

[14] Ma, W. & Mao, G. (2009). System structure partition based on graph theory and matrix theory. *J. Lanzhou Jiaotong Univ., 6*(1), 150-154.

[15] Niculescu, M., Irimia, C., Rosca, L., Grovu, M., & Guiman, M. (2017). Structural dynamic applications using principal component analysis. *CONAT 2016*, 90-99. https://doi.org/10.1007/978-3-319-45447-4_10

[16] Pozo, C., Ruiz-Femenia, R., Caballero, J., Guillen-Gosalbez, G., & Jimenez, L. (2012). On the use of Principal Component Analysis for reducing the number of environmental objectives in multi-objective optimization: Application to the design of chemical supply chains. *Chem. Eng. Sci., 69*(1), 146-158. https://doi.org/10.1016/j.ces.2011.10.018

[17] Qi, M., Fu, Z., Jing, Y., & Ma, Y. (2013). A Comprehensive Evaluation Method of Power Plant Units Based on Information Entropy and Principal Component Analysis. *Proceedings CSEE, 33*(2), 58-65. https://doi.org/10.13334/j.0258-8013.pcsee.2013.02.015

[18] Raul, P. J., Catherine, A.-P., & Stephan, A. (2018). Combining multi-objective optimization, principal component analysis and multiple criteria decision making for ecodesign of photovoltaic grid-connected systems. *Sustain. Energy Technol. Assessments, 27*(1), 94-101. https://doi.org/10.1016/j.seta.2018.03.008

[19] Sun, L. & Qian, W. (2009). Research on the comprehensive evaluation method based on principal component analysis. *Math. Pract. Theory., 18*(1), 15-20.
[20] Singarvel, B., Selvaraj, T., & Jeyapaul, R. (2014). Multi objective optimization in turning of EN25 Steel Using Taguchi Based Utility Concept Coupled with Principal Component Analysis. *Procedia Eng., 97*(1), 158-165. https://doi.org/10.1016/j.proeng.2014.12.237
[21] Sabio, N., Kostin, A., Guillen-Gosalbez, G., & Jimenez, L. (2012). Holistic minimization of the life cycle environmental impact of hydrogen infrastructures using multi-objective optimization and principal component analysis. *Int. J. Hydrogen Energy., 37*(6), 5385-5404. https://doi.org/10.1016/j.ijhydene.2011.09.039
[22] Sinha, P., Kumar, R., Singh, G. K., & Thomas, D. (2015). Multi-Objective Optimization of Wire EDM of AISI D3 Tool Steel Using Orthogonal Array with Principal Component Analysis. *Mater. Today Proc., 2*(1), 3778-3787. https://doi.org/10.1016/j.matpr.2015.07.183
[23] Vluymans, Ys., Cornelis, C., Herrera, F., & Saeys, Y. (2018). Multi-label classification using a fuzzy rough neighborhood consensus. *Inf. Sci. (Ny)., 433-434*(1), 96-114. https://doi.org/10.1016/j.ins.2017.12.034
[24] Wu, S., Gao, X., & B. M. (2003). *Data Warehousing and Data Mining*, first ed., Metallurgical Industry Press, Beijing.
[25] Wang, Z., Liu, J., Zhang, Q., Jiang, Y. (2015). Attribute reduction of the index system in circular economy based on rough sets. *Stat. Decis., 6*(1) 34-38.
[26] Wonggasem, K., Wagner, T., Trautmann, H., Biermann, D., & Weihs, C. (2013). Multi-objective Optimization of Hard Turning of AISI 6150 Using PCA-based Desirability Index for Correlated Objectives. *Procedia CIRP, 12*(1), 13-18. https://doi.org/10.1016/j.procir.2013.09.004
[27] Yi, S., Lai, Z., He, Z., Cheung, Y., & Liu, Y. (2017). Joint sparse principle component analysis. *Pattern Recognit., 61*(1), 524-536. https://doi.org/10.1016/j.patcog.2016.08.025
[28] Yu, Y., Pedrycz, W., & Miao, D. (2013). Neighborhood rough sets based multi-label classification for automatic image annotation. *Int. J. Approx. Reason., 54*(1), 1373-1387. https://doi.org/10.1016/j.ijar.2013.06.003
[29] Yu, Y., Miao, D., Zhang, Z., & Wang, L. (2013). Multi-label Classification Using Rough Sets. *Lect. Notes Artif. Intell.,* 119-126. https://doi.org/10.1007/978-3-642-41218-9_13
[30] Yu, Y., Pedrycz, W., Miao, D. & Zhang, H. (2013). Neighborhood Rough Sets Based Multi-label Classification. *Proc. 2013 Jt. IFSA World Congr. Nafips. Annu. Meet.,* 86-90. https://doi.org/10.1109/IFSA-NAFIPS.2013.6608380
[31] Yu, Y., Miao, D., Zhao, C., & Wang, Y. (2015). Knowledge Acquisition Methods for Multi-Label Decision System Based on Rough Sets. *J. Front. Comput. Sci. Technol., 9*(1), 94-104. https://doi.org/10.3778/j.issn.1673-9418.1406024
[32] Zhu, L. & Lv, B. (2004). Simple and convenient method of ISM. *Syst. Eng. Electron., 12*(1), 1815-1817.
[33] Zhang, Z., Zhao, T., & Chunhong, W. (2009). Application of the attribute reduction method based on rough set. *Sci. Technol. Manag. Res., 1*(1), 78-85.
[34] Zhang, L., Hu, Q., Zhou, Y., & Wang, X. (2014). Multi-label Attribute Evaluation Based on Fuzzy Rough Sets. *Lect. Notes Comput. Sci., 8536*(1), 100-108. https://doi.org/10.1007/978-3-319-08644-6_10
[35] Chatterjee, K., Pamucar, D., & Zavadskas, E. K. (2018). Evaluating the performance of suppliers based on using the R'AMATEL-MAIRCA method for green supply chain implementation in electronics industry. *Journal of Cleaner Production, 184*(1), 101-129. https://doi.org/10.1016/j.jclepro.2018.02.186
[36] Jagannath, R., Krishnedu, A., Samarji, K., & Dragan, P. (2018). A rough strength relational DEMATEL model for analysing the key success factors of hospital service quality. *Decision Making: Applications in Management and Engineering, 1*(1), 121-142. https://doi.org/10.31181/dmame1801121r
[37] Pamucar, D., Chatterjee, K., & Kazimieras Zavadskas, E. (2019).Assessment of third-party logistics provider using multi-criteria decision-making approach based on interval rough numbers. *Computers and Industrial Engineering, 127*(1), 383-407. https://doi.org/10.1016/j.cie.2018.10.023
[38] Stević, Ž., Đalić, I., Pamučar, D., Nunić, Z., Vesković, S., Vasiljević, M., & Tanackov, I. (2019). A new hybrid model for quality assessment of scientific conferences based on rough BWM and SERVQUAL. *Scientometrics, 119*(1), 1-30. https://doi.org/10.1007/s11192-019-03032-z
[39] Karavidic, Z. & Projovic, D. (2018). A multi-criteria decision-making (MCDM) model in the security forces operations based on rough sets. *Decision Making: Applications in Management and Engineering*, *1*(1), 97-120. https://doi.org/10.31181/dmame180197k

Contact information:

Ai WANG, PhD student
(Corresponding author)
University of Science and Technology Beijing,
No. 30 Xueyuan Road, Haidian District, Beijing, China
wangai22222@126.com

Xuedong GAO, Professor
University of Science and Technology Beijing,
No. 30 Xueyuan Road, Haidian District, Beijing, China
gaoxuedong@manage.ustb.edu.cn