Knowledge-Based Systems 205 (2020) 106290

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Meta-knowledge dictionary learning on 1-bit response data for student knowledge diagnosis

Yupei Zhang [a], Huan Dai [a], Yue Yun [a], Shuhui Liu [a], Andrew Lan [b], Xuequn Shang [a,*]

[a] School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[b] College of Information and Computer Sciences, University of Massachusetts Amherst, USA

ARTICLE INFO

Article history:
Received 21 March 2020
Received in revised form 14 June 2020
Accepted 17 July 2020
Available online 24 July 2020

Keywords:
Dictionary learning
Latent knowledge analysis
Student knowledge based-system
Cognitive diagnosis
Student clustering and classification

ABSTRACT

This paper focuses on the problem of student knowledge diagnosis that is a basic task of realizing personalized education. Most traditional methods rely on the question-concept matrix empirically designed by experts. However, the expert concepts are expensive and inter-overlapping in their constructions, leading to ambiguous explanations. With the intuition that each student can master a part of the knowledge involved in all questions, in this paper, we propose a novel learning-based model for student knowledge diagnosis, dubbed Meta-knowledge Dictionary Learning (metaDL). MetaDL aims to learn a meta-knowledge dictionary from student responses, where any knowledge entity (e.g., student, question or expert concept) is a linear combination of a few atoms in the meta-knowledge dictionary. The resultant problem could be effectively solved by developing the alternating direction method of multipliers. This study has three innovations: learning independent meta-knowledges instead of traditional complex concepts, sparely representing knowledge entity instead of densely weighted representation, and interpreting expert concepts with the resulting meta-knowledges. For evaluation, the diagnosis results from metaDL are used to group students and predict responses on two public datasets and a private dataset from our institution. The experiment results show that metaDL delivers an effective student knowledge diagnosis and then results in good performances on the two applications in comparison with other methods. This technique could provide significant insights into student's knowledge state and facilitate the progress on personalized education.

© 2020 Published by Elsevier B.V.

1. Introduction

Towards personalized education [1–3], student knowledge diagnosis (SKD), which aims to extract student's abilities, strengths, and interests, is a key problem to design different learning plans. Correctly analyzing student's knowledge status can reduce the unreasonable plans, where students may yield negative behaviors (e.g., dropout), and avoid the one-size-fits-all approach, where all students have the same plan. In the past two decades, many methods have been developed in educational theory-based research and computer-aided education.

The early SKD researches usually used cognitive diagnosis models from psychometric domain, such as Rasch model [4–6] and Item Response Theory (IRT) model [7,8]. Rasch model used a logic function to model the probability of an individual getting a correct response on a test item, by a trade-off between the individual's latent traits and the item difficulty. Due to the practical complication, IRT generalizes Rasch to consider more effects: discrimination in the 2-parameter logistic model [9] and guessing rate in the 3-parameter logistic model [10]. Chen C M et al. proposed a personalized e-learning system based on the IRT model and achieved effective performance on web-based market research [11]. However, both Rasch and IRT fail to provide the student's state of knowledge concepts so that they have limitations at personalized planning and remediation.

To place different plans for individuals or groups, diagnosing the skills whether they have been mastered becomes an important topic, *i.e.*, cognitive diagnosis. De La Torre et al. proposed the deterministic inputs, noisy "and" gate (DINA) model to identify the presence or absence of fine-grained skills for problem solving in a test, where the slipping and guessing factors were also considered [12,13]. Chen et al. integrated DINA and three online calibration methods for computerized adaptive testing [14]. In the work of Chen et al. [15], generalized DINA was used to explore the relationships among five reading comprehension skills defined by English language experts. In addition, DINA is a conjunctive model where all skills of an item are all mastered for a correct answer, while an incorrect response to an item might be due to

* Corresponding author.
E-mail addresses: ypzhaang@nwpu.edu.cn (Y. Zhang), daihuan@mail.nwpu.edu.cn (H. Dai), yundayue@mail.nwpu.edu.cn (Y. Yun), lsh@mail.nwpu.edu.cn (S. Liu), andrewlan@cs.umass.edu (A. Lan), shang@nwpu.edu.cn (X. Shang).

https://doi.org/10.1016/j.knosys.2020.106290
0950-7051/© 2020 Published by Elsevier B.V.

some misconceptions. Bor et al. proposed a disjunctive model, Bug-DINO, under the assumption that students are expected to correctly answer an item only if they do not possess any of the misconceptions on the item [16]. Based on both ideas, Dimitrov et al. proposed the least-squares distance model to manage the disjunctive assumption and the conjunctive assumption [17].

In the above cognitive diagnosis models, Q-matrix represents item-attribute relationships such that the responses to the items could reflect attribute configurations of students. Therefore, Q-matrix is the core component that determines the quality of a cognitive diagnosis. However, the Q-matrix in previous works was manually designed by experts with subjective experience. The manual Q-matrix has the following shortcomings: (1) the skills or concepts in Q-matrix are often overlapped, reducing the specificity on individuals; (2) the incompleteness is usually suffered because of the complex combination of expert concepts; (3) subjective experience often causes an attention bias to difficult skills or concepts. Several studies have made attempts on designing a Q-matrix based on the given data for cognitive diagnosis. Liu et al. designed an extra T-matrix that connects the observed response distribution and the model structure to identify a Q-matrix [18]. Chiu et al. develops a Q-matrix refinement method based on the nonparametric classification model, which enumerates all possible patterns of existing knowledge points and students' possible responses and then calculates the distances between observations and ideal responses, to obtain a reliable Q matrix [19]. Kohn et al. built the complete Q-matrix to guarantee the identifiability of all possible proficiency classes among individuals [20]. However, these designed Q-matrices are data-agnostic and thus the concepts are not optimal for a specific problem.

To learn a problem-specific Q-matrix, Lan et al. proposed a sparse factor model for learning analytics (SPARFA) to estimate the student's knowledge of concepts from response data. SPARFA learns several abstract complex concepts to diagnosis the mastery status, followed by grouping students and labeling the question's difficulty [21]. However, it is limited to learning a small number of concepts such that the learned concepts are too overlapped to distinguish items and students. Besides, SPARFA fails to tackle a new student case. In this study, we propose a novel method to learn a meta-knowledge dictionary, where the concepts in the dictionary are overcomplete. As shown in Fig. 1, our method is based on the intuition that every student (e.g., L_1, L_2, ...) could grasp a subset of the meta-knowledges (e.g., k_1, k_2, ...) involved in the given questions, while different students usually grasp different subsets. With this assumption, student or other knowledge entities could be represented as a sparse combination of these meta-knowledges. The purpose of our method is to learn a set of meta-knowledges and the representations of students over the meta-knowledges. Our major contributions lie in:

- A new knowledge representation method. We propose a novel abstract concept for student knowledge diagnosis, which is fine-grained and inseparable knowledge elements, dubbed meta-knowledge. While traditional methods learn or design comprehensive concepts, our method sparsely represents each knowledge entity in an overcomplete space.
- A meta-knowledge learning model. We propose a meta-knowledge dictionary learning method (metaDL) to learn the latent concepts from students' binary responses. Besides, the guessing and slipping factors are integrated by the noise in metaDL. MetaDL is the first attempt on building a gene dictionary for the specific response data.
- A 1-bit dictionary-learning algorithm. We develop the alternating direction method of multipliers (ADMM) to tackle the objective problem of metaDL. Note that metaDL is learning on the 1-bit data indicating correct or incorrect responses to questions.

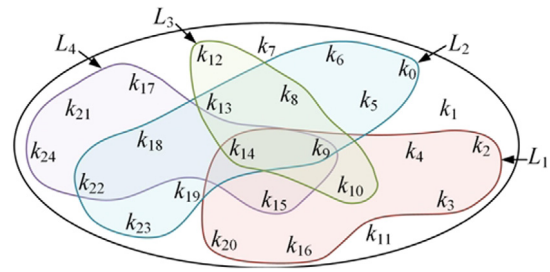


Fig. 1. The illustration of meta-knowledges. L_1, L_2, L_3 and L_4 indicates four students. k_1, k_2,..., k_{20} represents the meta-knowledges. Different subsets of meta-knowledge highlighted by different colors shows different student knowledge structures. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- A explanation with expert concepts. We employed a sparse linear regression to build the connection between the meta-knowledges and the Q-matrix. As a result, the explanation of student knowledge diagnosis can be achieved on expert concepts, if the Q-matrix is available.
- A real-world experiment. We apply metaDL to analysis students in our course of C language programming of an international class. With our questionnaires, it is found that student who favors mathematics and physics could achieve a higher score in the C language learning. Besides, two publicly available datasets are used to validate metaDL on student grouping and response prediction.

In addition, metaDL could be used without Q-matrix for various tasks in education, such as course selection and question recommendation [22]. The rest of this paper is organized as follows. In Section 2, we introduce the related work including cognitive diagnosis and sparse dictionary learning. In Section 3, we reformulate the SKD problem in mathematics and present the metaDL method, followed by the learning algorithm of metaDL. Experiment results on two public datasets and a real-course data are exhibited in Section 4. Section 5 exhibits discussions and conclusions.

2. Related work

This section briefly introduces the mainly related works that support our study on the problem of student knowledge diagnosis, including cognitive diagnosis [21,23] and sparse dictionary learning [24–26].

2.1. Cognitive diagnosis

In the educational psychology domain, cognitive diagnosis techniques are usually used to investigate student knowledge mastery and predict student behaviors by tracking students learning performance [3]. The purpose of cognitive diagnosis is to establish the relationship between the students and the given expert concepts according to the observed responses. There are two common psychometrical models in existing studies, i.e., item response theory-based model (IRT) and deterministic inputs, noisy-and gate model (DINA). IRT models the probability of a correct response to an item (i.e., question) as a S-shape function of student's ability and item's difficult. IRT can be formalized as:

$$P(\theta) = c + (1 - c)\frac{1}{1 + e^{-a(\theta - b)}} \tag{1}$$

where $P(\theta)$ is the probability of a correct response; θ and b indicate student's ability and item's difficult respectively; a is a scale parameter; and c is to model the guessing. Eq. (1) shows

the IRT model of three parameters [23], which degenerates into the two-parameter IRT by removing c and degenerates into the one-parameter IRT by removing c and a. Note that the one-parameter IRT is also referred to as Rasch model [27]. While IRT-based models often require a predefined Q-matrix to show the concepts per question, learning concepts from observations becomes an interesting topic. Deep learning-based method has been adopted for knowledge diagnosis [28] in a large amount of observations. However, there is usually a small number of students in a real-world classroom.

In traditional educational case, sparse factor analysis for learning and content analytics (SPARFA) [21] showed a good performance on latent knowledge concept learning. SPARFA models the question's response as:

$$\begin{aligned} Y_{i,j} &\sim \text{Ber}\left(\Phi\left(Z_{i,j}\right)\right), \\ Z_{i,j} &= \mathbf{w}_i^T \mathbf{c}_j + \mu_i, \qquad \forall i, j \end{aligned} \tag{2}$$

where $Y_{i,j} \in \{0, 1\}$ for the student j on the question i, with 1 being a correct response and 0 being an incorrect response; $\text{Ber}(x)$ is a Bernoulli distribution with the variable x; $\Phi(z)$ is an inverse link function that maps a real z to a random binary variable; $\mathbf{c}_j$ is the weight representation of the student j on all concepts; $\mathbf{w}_i$ is the question-concept association of the question i; μ_i models the intrinsic difficulty of the question i. With the learned latent concepts, SPARFA could not only deliver the question's difficulty for question selecting, but also provide a strong explanatory power with the learned factors. However, SPARFA is variation of matrix factorization which is difficult to tackle a new student case. On the other hand, SPARFA is limited to learning a few abstract concepts so that the diagnosis result cannot distinguish the involved students.

2.2. Sparse dictionary learning

Sparse dictionary learning (SDL) [29,30] aims to learn a suitable dictionary from densely expressed samples and then convert the samples into sparse representation, thereby simplifying the learning task and reducing the model complexity. The dictionary is usually assumed to be an overcomplete bases composed of independent dictionary atoms. Let $\mathbf{D}$ be the dictionary, $\mathbf{x}_i$ be the objective signal and $\mathbf{r}_i$ be the sparse representation of $\mathbf{x}_i$ on $\mathbf{D}$. SDL learns $\mathbf{D}$ and $\mathbf{r}_i$ from the given data $\mathbf{x}_i$ and could be formulated as:

$$\underset{\mathbf{D}\in\Omega, r_i\in R^n}{\arg\min} \sum_{i=1}^{K} \|x_i - \mathbf{D}r_i\|_2^2 + \lambda\|r_i\|_0 \tag{3}$$

where K is the number of students. Since L_0-norm is difficult to manage, it is commonly replaced by the convex L_1-norm [31].

SDL has attracted many successes in compressed sensing [32,33] and machine learning [34,35] in the past decade. This study considers SDL for student knowledge diagnosis is due to the following advantages: (1) The explanation. SDL has a direct connection from the representation and the dictionary thanks to the linear maps in Eq. (3). (2) The sparsity. SDL allocates a few atoms to each student $\mathbf{x}_i$ so as to increase the discrimination between student's knowledge structures. (3) The small-sample-size issue. In traditional class, the data is usually small and insufficient for learning a complex model, e.g., deep neural network [36]. Another key consideration is that this study assumes each student can grasp a part of the knowledges involved in an exam and each question as well contains a part of these knowledges.

However, traditional SDL models are usually focused on dense data rather than binary data. Fig. 2(a) shows the data formulation in our problem. Although recent studies [37] have mentioned dictionary models on binary data, those models fail to consider the specific education problem. To this end, this study proposes a new binary dictionary leaning technique for the SKD problem.

3. Method

In this section, we explicitly formalize the SKD problem in mathematics. To manage this problem, we propose a new learning-based method termed Meta-knowledge Dictionary Learning (metaDL), inspired by the knowledge concept [21] and sparse dictionary learning [37].

3.1. Student knowledge diagnosis problem

Student knowledge diagnosis aims to analysis student's knowledge mastery, where the key point is the knowledge representation. While a Q-matrix is often created on experts' experience, learning knowledge points from data is a more promising way. In general, the response to a question is determined by the student's mastery state on all knowledge points involved in questions. That is, the response could be caused by *student's knowledge vector* and *question's knowledge vector*. Rather than empirically defining knowledge points for knowledge vector, this study uses a fine-grained knowledge concept, i.e., meta-knowledge, for SKD.

The meta-knowledges can be considered as knowledge cells that are dependent and inseparable. As shown in Fig. 1, considering an exam paper of a meta-knowledge set, every student masters a subset of this meta-knowledge set, while different student grasps different meta-knowledge subsets. Furthermore, the meta-knowledge that only belongs to a student is called specific meta-knowledge, e.g., k_2 for L_1, while belongs to multiple students is called common meta-knowledge, e.g., k_9 for the four students. The set of all meta-knowledges involved in this exam paper are called knowledge dictionary. With these definitions, we have the following assumption:

Assumption 1. Given a dataset of N students and Q questions, let $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_i, \ldots, \mathbf{d}_k] \in R^{Q\times K}$ be the knowledge dictionary where K is the number of meta-knowledges, shown in Fig. 2. The jth student's mastery knowledge can be represented as $\mathbf{x}_j \in R^{K\times 1}$. As a result, the response y_{ij} to the ith question can be:

$$y_{ij} = \mathbf{D}_i\mathbf{x}_j \tag{4}$$

where $\mathbf{D}_i$ is a row vector of $\mathbf{D}$ corresponding to the ith question.

However, in practical education cases, students often suffer from slipping or guessing. To model this information, we introduce an additional term w_{ij} for the student i on the question j into Eq. (4). Overall, this study models the problem of student knowledge diagnosis as follows:

Problem 1. Consider that N students answer Q questions leading to a response data $\mathbf{Y}$. $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n] \in R^{Q\times N}$, with $\mathbf{y}_i$ being the ith student's responses to the Q questions. With Assumption 1, this study aims to learn the meta-knowledge dictionary $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_k] \in R^{Q\times K}$, the students' knowledge mastery matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in R^{K\times N}$, and the matrix $\mathbf{W} \in R^{Q\times N}$ for the slipping and guessing.

3.2. The proposed model

To achieve a solution to Problem 1, we propose a sparse dictionary learning model, dubbed meta-knowledge dictionary leaning (metaDL). The workflow of metaDL is illustrated in Fig. 3. MetaDL is formulated as:

$$\mathbf{Y} = \text{sign}(\mathbf{DX} + \mathbf{W}) \tag{5}$$

where $y_{ij} \in \{0, 1\}$ is the response of the student j to the question i, where 1 indicates a correct response while 0 indicates an incorrect response. $\text{Sign}(\cdot)$ in Eq. (5) is the signum function. With Assumption 1 and educational priors, we have the following observations:

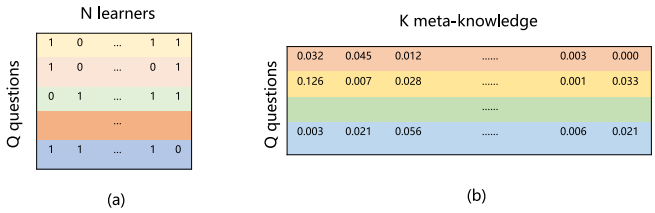


Fig. 2. The data formulation in this study, including (a) the question-student responses for inputs, **Y** and (b) the meta-knowledge dictionary learned, **D**. The elements 0s and 1s corresponds to correct and incorrect response in (a), respectively. The weights in (b) is the corresponding importance between meta-knowledges and questions.

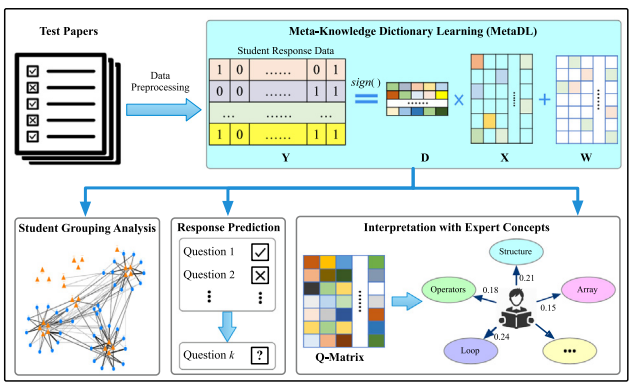


Fig. 3. An overview of this study with the proposed MetaDL method.

- Meta-knowledges dictionary **D** are usually overcomplete with respective to each student and each question. The reason is that an exam paper usually contains more meta-knowledges than that involved by a student or a question.
- Each student usually grasps a subset of the meta-dictionary **D**. Thus, $\mathbf{x}_i$ is a sparse vector for choosing a few meta-knowledges from **D**.
- The slipping and guessing usually appear on a few questions. Thus, the matrix **W** is believed to be sparse for noise modeling.

With above assumptions and observations, the objective problem of our metaDL is cast as follows:

$$\begin{aligned} &\underset{\mathbf{D},\mathbf{X},\mathbf{W}}{\text{minimize}} \sum_{i,j} \left\| y_{ij} - \text{sign}(\mathbf{D}_i\mathbf{x}_j + w_{ij}) \right\|_F^2 \\ &\text{subject to } \|\mathbf{D}_i\|_2 \le 1,\ \left\|\mathbf{x}_j\right\|_0 \le s,\ \|\mathbf{W}\|_0 \le k \end{aligned} \tag{6}$$

where s is the sparsity of $\mathbf{x}_j$ for student's knowledge subset; k is the sparsity of **W** for slipping and guessing.

3.3. Meta-knowledge dictionary learning

Unfortunately, the proposed objective problem (6) is a non-convex problem due to the signum function and the L_0 pseudo-norm constrains. To achieve an optimization solution of problem (6), we relax L_0-norm into L_1-norm as in [21], and use the continuous function $\sigma(x) = \frac{1}{1+e^{-x}}$ to approximate sign(x). Based on experimental practice, we really use the function $\sigma_\alpha(x) = \frac{1}{1+e^{-\alpha x}}$ to get an approximation. Then, the objective problem is reduced into an unconstraint problem as:

$$\begin{aligned} &\underset{\mathrm{D,X,W}}{\text{minimize}} \|\mathbf{Y} - \sigma_\alpha(\mathbf{DX} + \mathbf{W})\|_F^2 + \lambda \sum_{j=1}^{N} \left\|\mathbf{x}_j\right\|_1 \\ &\qquad + \beta\|\mathbf{W}\|_1 + \gamma\left(1 - \sum_{i=1}^{Q} \|D_i\|_2\right) \end{aligned} \tag{7}$$

where $\lambda > 0$, $\beta > 0$ are the trade-off parameters for sparsity level; γ is a penalty parameter and the scale factor α in $\sigma(x)$ is set to 10 in this study.

Like most dictionary learning algorithms [38–40], we here develop the alternating direction method of multipliers (ADMM) [41] to solve problem (7). To make the objective function separable, we introduce an auxiliary variable **J**, and rewritten (7) as:

$$\begin{aligned} &\underset{\mathrm{DX,W}}{\text{minimize}} \|\mathbf{Y} - \sigma_\alpha(\mathbf{DX} + \mathbf{W})\|_F^2 + \lambda \sum_{j=1}^{N} \left\|\mathbf{x}_j\right\|_1 \\ &\qquad + \beta\|\mathbf{W}\|_1 + \gamma\left(1 - \sum_{i=1}^{Q} \|D_i\|_2\right) \\ &\text{s.t. } \mathbf{W} = \mathbf{J} \end{aligned} \tag{8}$$

Then, the augmented Lagrangian function of problem (8) can be written into:

$$\begin{aligned} \mathcal{L}(\mathbf{D},\mathbf{X},\mathbf{W},\mathbf{J},M) &= \|\mathbf{Y} - \sigma_\alpha(\mathbf{DX} + \mathbf{W})\|_F^2 + \lambda \sum_{j=1}^{N} \left\|\mathbf{x}_j\right\|_1 \\ &+ \beta\|\mathbf{J}\|_1 + < M, \mathbf{W} - \mathbf{J} > + \frac{\mu}{2}\|\mathbf{W} - \mathbf{J}\|_F^2 \\ &+ \gamma\left(1 - \sum_{i=1}^{Q} \|D_i\|_2\right) \end{aligned} \tag{9}$$

where M is Lagrange multipliers and $\mu > 0$ is a penalty parameter. According to ADMM, the variables are updated alternately by minimizing their corresponding subproblems with fixing other variables. We provide the main steps of our learning algorithm as follows.

(1) Updating **J** is equivalent to minimizing the following objective function:

$$\begin{aligned} \mathcal{L}_1 &= \beta\|\mathbf{J}\|_1 + < M_k, \mathbf{W}_k - \mathbf{J} > + \frac{\mu}{2}\|\mathbf{W}_k - \mathbf{J}\|_F^2 \\ &= \beta\|\mathbf{J}\|_1 + \frac{\mu}{2}\left\|\mathbf{J} - \left(\mathbf{W}_k + \frac{1}{\mu}M_k\right)\right\|_F^2 \end{aligned} \tag{10}$$

which was developed into a closed-form solution in [42], shown as follows.

$$\mathbf{J}_{k+1} = \max\left\{S_{\frac{\beta}{\mu}}\left(\mathbf{W}_k + \frac{1}{\mu}M_k\right), 0\right\}. \tag{11}$$

where $S_{\frac{\beta}{\mu}}$ indicates the single value thresholding operator.

(2) Updating **W** aims to solve the following problem:

$$\begin{aligned} \mathcal{L} &= \min_{\mathbf{W}} \|\mathbf{Y} - \sigma_\alpha(\mathbf{DX} + \mathbf{W})\|_F^2 + < M_k, \mathbf{W} - \mathbf{J}_k > \\ &\quad + \frac{\mu}{2}\|\mathbf{W} - \mathbf{J}_k\|_F^2 \\ &= \min_{\mathbf{W}} \|\mathbf{Y} - \sigma_\alpha(\mathbf{DX} + \mathbf{W})\|_F^2 + \frac{\mu}{2}\left\|\mathbf{W} - (\mathbf{J}_k + \frac{1}{\mu}M_k)\right\|_F^2 \end{aligned} \tag{12}$$

We adopt the commonly used gradient descent algorithm and update the matrix **W** by:

$$\mathbf{W}_{k+1} = \mathbf{W}_k - \alpha\nabla_{\mathbf{W}}\mathcal{L}$$

where the partial derivative $\nabla_{\mathbf{W}}\mathcal{L}$ is equal to

$$2\left[\mathbf{Y}-\sigma_\alpha(\mathbf{DX}+\mathbf{W})\right]\nabla_{\mathbf{W}}\sigma_\alpha(\mathbf{DX}+\mathbf{W}) +\mu\left[\mathbf{W}-\left(\mathbf{J}_k+\tfrac{1}{\mu}M_k\right)\right] \tag{13}$$

where $\nabla_{\mathbf{W}}$ is to obtain the gradient of $\sigma_\alpha(\mathbf{DX}+\mathbf{W})$ with respect to $\mathbf{W}$.

(3) Updating $\mathbf{X}$ and $\mathbf{D}$ is to solve the following problem:

$$\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y}-\sigma_\alpha(\mathbf{DX}+\mathbf{W})\|_F^2+\lambda\sum_{j=1}^{M}\left\|\mathbf{x}_j\right\|_1 + \gamma\left(1-\sum_{i=1}^{Q}\|\mathbf{D}_i\|_2\right) \tag{14}$$

As most dictionary learning algorithms, we employ the coordinate decent strategy to iteratively optimize $\mathbf{X}$ and $\mathbf{D}$ as follows.

- Fixing $\mathbf{D}$, updating $\mathbf{X}$ can be decomposed into optimize every $\mathbf{x}_i$ independently. We optimize $\mathbf{x}_i$ using the BIHT algorithm:

$$\mathbf{x}_j=\mathrm{BIHT}(\mathbf{y}_j,\mathbf{D})\ 1\le j\le M \tag{15}$$

  which has a closed solution referred to the work [37].
- Fixing $\mathbf{X}$, updating $\mathbf{D}$ is usually implemented atom by atom [43]. Here, we propose an new route based on the work [37]. The problem can be rewritten into:

$$\min_{\mathbf{D}_i}\sum_{i=1}^{N}\sum_{j=1}^{Q}\left|y_{ij}-\sigma_\alpha\left(\mathbf{D}_i\mathbf{x}_j+w_{ij}\right)\right|^2 + \gamma\left(1-\sum_{i=1}^{Q}\|\mathbf{D}_i\|_2\right) \tag{16}$$

where $\mathbf{D}_i$ is the ith dictionary row and $\mathbf{x}_j$ is the jth column vector of matrix $\mathbf{X}$. w_{ij} is an element of the bias matrix $\mathbf{W}$. Therefore, the optimization problem in Eq. (16) can be divided into Q sub-optimization problems to update the dictionary matrix $\mathbf{D}$ row by row. Each sub-problem can be written into:

$$\min_{\mathbf{D}_i}\mathcal{F}(\mathbf{D}_i)=\sum_{i=1}^{N}\left|y_{ij}-\sigma_\alpha(\mathbf{D}_i\mathbf{x}_j+w_{ij})\right|^2 + \frac{\gamma}{2}(1-\|\mathbf{D}_i\|_2)1\le j\le N \tag{17}$$

The cost function of Eq. (17) is a continuous function and can be minimized by gradient decent [37]. Wherein, the gradient step is:

$$\mathbf{D}_{k+1}=\mathbf{D}_k-\beta\nabla_{\mathbf{D}_k}\mathcal{F} \tag{18}$$

where the partial derivative $\nabla_{\mathbf{D}_k}\mathcal{F}$ is equal to

$$-\sum_{i=1}^{N}2[y_{ij}-\sigma_\alpha(\mathbf{D}_i\mathbf{x}_j+w_{ij})]\nabla_{\mathbf{D}_i}\sigma_\alpha(\mathbf{D}_i\mathbf{x}_j+w_{ij})+\gamma\mathbf{D}_i \tag{19}$$

where $\nabla_{\mathbf{D}_i}$ is the gradient of $\sigma_\alpha(\mathbf{D}_i\mathbf{x}_j+w_{ij})$ with respect to $\mathbf{D}_i$. The procedure of solving the proposed metaDL problem is briefly described in Algorithm 1.

3.4. Knowledge diagnosis with expert concepts

MetaDL aims to learn $\mathbf{D}$, $\mathbf{X}$ and $\mathbf{W}$ from the given binary responses $\mathbf{Y}$, where the meta-knowledges $\mathbf{D}$ are estimated from data rather than experts. Enabling meta-knowledges to be interpreted on expert concepts when the expert Q-matrix is given could show a more useful result in the real-world classroom. Since we assume expert concepts can be linearly combined by subsets of meta-knowledges, we use a linear regression to build the relationships between meta-knowledges and expert concepts.

Algorithm 1 Meta-knowledge Dictionary Learning

Input: Response data $\mathbf{Y}$
Parameter: $\mathbf{D}=\mathbf{0}$, $\mathbf{X}=\mathbf{D}^{-1}\mathbf{Y}$, $\mathbf{W}=\mathbf{0}$, $\lambda=0.03$, $\beta=0.4$, $\gamma=0.01$, $\alpha=10$
Output: Meta-dictionary $\mathbf{D}$

while not converged ($k=1,0,\cdots$) do
  fix the others and update $\mathbf{J}$ by:
  $\mathbf{J}_{k+1}=\max\left\{\mathrm{S}_{\frac{\beta}{\mu}}\left(\mathbf{W}_k+\frac{1}{\mu}M_k\right),0\right\}$
  fix the others and update $\mathbf{W}$ by:
  $\mathbf{W}_{k+1}=\mathbf{W}_k-\alpha\nabla_{\mathbf{W}}\mathcal{L}$
  fix the others and update $\mathbf{X}$ by:
  $\mathbf{x}_j=\mathrm{BIHT}(\mathbf{y}_j,\mathbf{D})=\mathrm{H}_M(\mathbf{x}_k+\mathbf{D}^T(\mathbf{y}_j-\mathrm{sigmoid}(\mathbf{Dx}_k+\mathbf{W})))$
  fix the others and update $\mathbf{D}$ by:
  $\mathbf{D}_{k+1}=\mathbf{D}_k-\beta\nabla_{\mathbf{D}_k}\mathcal{F}$
  update the multipliers:
  $M_{k+1}=M_k+\mu_k(\mathbf{W}_{k+1}-\mathbf{J}_{k+1})$
  where μ is a constant value.
  check convergence:
  $\frac{\|\mathbf{Y}-\mathrm{sigmoid}(\mathbf{DX}+\mathbf{W})\|_F^2}{\|\mathbf{Y}\|_F^2}<\varepsilon$
End while.

Supposed that the given Q-matrix $\mathbf{T}\in R^{N\times P}$ is composed of N questions associated with P concepts. Note that we set $T_{ip}=1$ if the concept p is contained in the question i, otherwise $T_{ip}=0$. Based on our assumption of meta-knowledges, the expert concept $\mathbf{T}_p$ can be caused by $\mathbf{T}_p=\mathbf{Da}_p$, where $\mathbf{a}_p$ represents the associations between meta-knowledges and the expert concept. That is, $\mathbf{T}=\mathbf{DA}$ where the pth column of $\mathbf{A}$ is $\mathbf{a}_p$. Due to the fact that each expert concept is associated with a few meta-knowledges in $\mathbf{D}$, we achieve $\mathbf{A}$ by minimizing:

$$\|\mathbf{T}-\mathbf{DA}\|_{\mathrm{F}}^2+\eta\sum_{p=1}^{P}\left\|\mathbf{a}_p\right\|_1 \tag{20}$$

which is exactly the problem of lasso regression [44]. Hence, the ith student can be represented on $\mathbf{D}$ as $\mathbf{x}_i$; all expert concepts can be represented on $\mathbf{D}$ as $\mathbf{A}$. As a consequence, we could achieve the state of the ith student on expert concepts by $\mathbf{u}_i=\mathbf{x}_i^T\mathbf{A}$, where u_{ip} indicates the binary state of the ith student on the pth expert concept. With such analysis results, the teacher could place a personalized plan to each student.

4. Experiments

In this section, we validate metaDL on two public datasets and a private real-world educational dataset. In general, two fundamental tasks in education are finding the students with a same knowledge structure and predicting a student's performance on new questions. Therefore, we conduct the two real-world concerned tasks, i.e., student grouping and student response prediction. To convince our method, we compare with Original that uses original features, SPARFA that learns abstract concepts and BOBCS that is a one-bit dictionary learning method. All experiment codes are implemented by MATLAB R2018a. Our codes and data have been made available on our website.[1]

[1] https://github.com/ypzhaang/Student-Knowledge-Diagnosis.

Table 1
Evaluation results of student grouping with the four methods on FrsCub.

	$c = 2$			$c = 3$			$c = 4$			$c = 5$			$c = 6$			$c = 7$			$c = 8$		
	CP	DVI	Ratio	CP	DVI	Ratio	CP	DVI	Ratio	CP	DVI	Ratio	CP	DVI	Ratio	CP	DVI	Ratio	CP	DVI	Ratio
Original	3.556	0.323	11.008	4.290	0.445	9.640	4.906	0.521	9.416	1.909	0.232	8.230	3.312	0.431	7.685	3.015	0.403	7.481	2.223	0.317	7.014
BOBCS(500 features)	0.050	0.321	0.156	0.038	0.357	0.106	0.022	0.250	0.089	0.043	0.370	0.117	0.023	0.370	0.061	0.032	0.369	0.087	0.020	0.274	0.071
SPARFA(5 features)	33.245	0.062	536.208	17.264	0.082	210.537	19.528	0.074	263.893	16.454	0.086	191.326	12.553	0.075	167.372	17.203	0.088	195.484	8.138	0.093	87.507
MetaDL(500 features)	**0.038**	**0.424**	**0.090**	**0.036**	**0.484**	**0.073**	**0.023**	**0.528**	**0.043**	**0.023**	**0.393**	**0.06**	**0.027**	**0.517**	**0.053**	**0.030**	**0.414**	**0.073**	**0.028**	**0.312**	**0.090**

4.1. Datasets

(1) **FrcSub**: This dataset is produced by the middle school students on the problem of fraction subtraction [3]. There are two elements: 1 indicates a correct response and 0 indicates an incorrect response. This dataset contains 536 students and 20 questions. There is no missing data.

(2) **ITI**: This dataset is publicly available at https://sites.google.com/site/assistmentsdata/projects/kansas-project. It contains the responses of 2846 students to 25 individual items in the Integer Test Items file, where 1 is for correct response and 0 is for incorrect response. The students are from several grades, where 8% of the students are from grade 6, 49% in grade 7, 39% in grade 8 and 4% in grade 10. We remove the missed responses and obtain 2774 students' responses.

(3) **NPU-C**: This private dataset was collected from international students who enrolled the C Language Program final exam in December 2018 in our institution. We collected the responses of 39 students to 20 questions. The dataset is also formed into binary elements, where 0 is for incorrect response and 1 is for correct response. To collect the background information, we made a questionnaire with twenty questions including gender, age, English and Math grade of College Entrance Examination, their favorite subjects, and so on. Note that there is no missing cases in our data. In addition, we constructed the Q matrix $\mathbf{T}$ from C language experts. The matrix $\mathbf{T}$ shows the correlations between 20 questions and 25 expert concepts.

4.2. Comparison method

In experiments, we compare our method with the three methods that is performed as follows:

Original: Original is to use the original data for the two tasks. That is, the response matrix is considered by using questions as features to cluster students or predict student responses. This comparison is to survey the effects of feature leaning on distinguishing students.

SPRAFA: SPARFA [21] is also evaluated by learning a few latent abstract concepts. Based on the learned concepts from these datasets, the two designed experiments are carried out. This comparison aims to have insights into our innovation of meta-knowledge learning.

BOBCS: BOBCS [37], a state-of-the-art one-bit compressed sensing algorithm, is adopted to learn the dictionary from responses. Then, the designed experiments are conducted on the learned dictionary features. This comparison aims to convince our innovation on binary dictionary leaning.

4.3. Student grouping

In student knowledge diagnosis, student grouping is a basic problem for educational intervention, such as making learning plan and recommending learning contexts for different students. In this experiment, we test different methods on the three datasets. Then the learned student representations are used for grouping student with the k-means algorithm. Since students whose meta-knowledges overlap much each other should be clustered together, students in a cluster master the same knowledge concepts. As a result, educational intervention could be made accordingly. Here, we evaluate the results in terms of Dunn Validity Index (DVI) [45] and Compactness Index (CP) [46], as follows:

$$\mathrm{DVI}=\frac{\min\limits_{0<m\neq n<k}\left\{\min\limits_{\forall x_i\in\Omega_m,\forall x_j\in\Omega_n}\left\|x_i-x_j\right\|\right\}}{\max\limits_{0<m\leq k}\left\{\max\limits_{\forall x_i,\forall x_j\in\Omega_m}\left\|x_i-x_j\right\|\right\}} \tag{21}$$

where max{} and min{} are to obtain the greatest element and the smallest element from a set, respectively; x_i and x_j are two students in consideration. The bigger DVI indicates a better cluster that has a low intra-class distance and a high inter-class distance.

$$\mathrm{CP}=\frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{|\Omega_i|}\sum_{x_i\in\Omega_i}\|x_i-w_i\|\right) \tag{22}$$

where k is the number of cluster and w_i represents a cluster center. The smaller CP delivers the better cluster that has a low intra-class average distance. Furthermore, we use the ratio of CP to DVI to measure the clustering quality as:

$$\mathrm{Ratio}=\frac{\mathrm{CP}}{\mathrm{DVI}} \tag{23}$$

As mentioned above, the smaller the ratio is, the better performance of a clustering method. To interpret the cluster results, we define Consistency Degree (CD) on the student responses in each cluster as:

$$\mathrm{CD}=\frac{\#\{\text{Commonly Correct or Incorrect Responses}\}}{\#\{\text{Questions}\}} \tag{24}$$

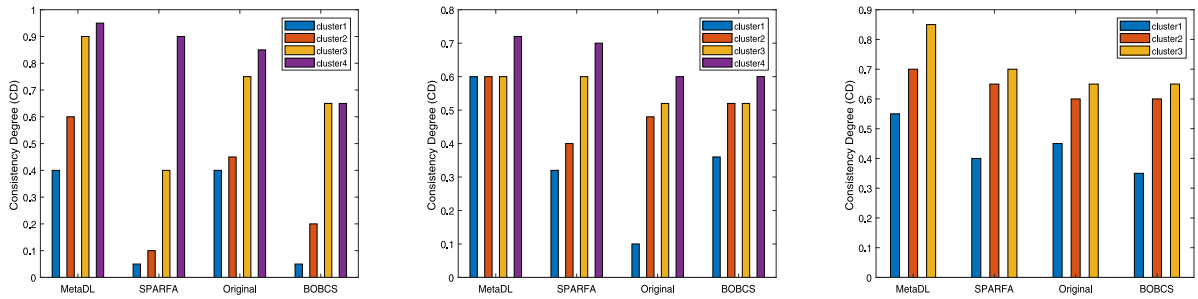
where #{} is to count the number of elements in a set. CD measures whether the students with the same knowledge structure are grouped into a cluster. The maximum CD value is 1, meaning that all students in the cluster have same responses to all questions. While, CD = 0 meaning that all students have no any same responses to all questions. Therefore, a cluster result is better when its CD is close to 1.

4.3.1. FrcSub

We vary the cluster number $c = \{2, 3, 4, 5, 6, 7, 8\}$ to select the one with the lowest ratio and report all evaluation results in Table 1. Because of learning more features, BOBCS and MetaDL with 500 features get a smaller CP value and a larger DVI value than Original and SPARFA. The ratio of metaDL is better than other methods, and achieves the best performance when cluster number is 4. Note that SPARFA leads to the large ratio, since it results in a large CP against a small DVI. The reason is that SPARFA aims to learn a small number of "abstract concepts" and thus make student representations much overlapped. Furthermore, metaDL considers the slipping and guessing so as to deliver better performance than BOBCS. Besides, more knowledge points deliver more discriminative student representations.

Based on the clustering results of MetaDL with 4 clusters, we seek the questions that most of these students simultaneously give correct or incorrect answers and then present the No. of these questions in Table 2. Table 2 shows that all students in cluster No. 1 are good at questions 6 and 14, but weak at the questions 16, 17, 19 and 20. And there have same explanations for cluster No. 2 and cluster No. 3. On the contrary, metaDL finds more commonly weak questions for cluster No. 4. According to these results, our algorithm could automatically suggest different remediations for different student groups.

Fig. 4(a) shows the student grouping results on FrcSub in terms of CD. As is shown, metaDL yields the biggest CD in every cluster than other compared methods.



(a) (b) (c)

Fig. 4. The student grouping results on (a) FrcSub, (b) ITI and (c) NPU-C in terms of CD, using the four mentioned methods.

Table 2
Question analysis of student groups from MetaDL on FrcSub.

Cluster	No. of correct questions	No. of incorrect questions
1	6 14	16 17 19 20
2	1 2 4 6 8 11 12 14 16	17 19
3	1 2 7 9 10 13 14 15 17 19 20	16
4	–	1 14 17 19 20

Table 3
Question analysis of student groups from MetaDL on ITI.

Cluster	No. of correct questions	No. of incorrect questions
1	1 6 10 11 15 17 18 19 20	–
2	2 6 9 10 11 15 17 20	1 18
3	6 10 11 12 13 15 17 18 20	1
4	1 2 6 10 11 12 13 15 17 18 19 20	4 16

4.3.2. ITI

Since this data is from four grades, we group these students into four clusters with all mentioned algorithms. The results on ITI in terms of CD are shown in Fig. 4(b). Fig. 4(b) shows that metaDL achieves bigger CDs in all clusters than other methods. Table 3 shows the No. of the commonly correct or incorrect questions in each cluster. The students in cluster 1 are good at No. 1, 6, 10, 11, 15, 17, 18, 19, 20 while there is no commonly incorrect questions. The students in cluster 2 are good at questions 2, 6, …, 20, while are bad at questions 1, 18. The results provide specific suggestions for remedy training based on the common performance.

4.3.3. NPU-C

We test the four methods varying the number of clusters $c = \{2, 3, 4, 5\}$ and then show CP, DVI and the Ratio in Table 4. As is shown, the more features the methods learn, the better performance the method achieves. Among the four mentioned methods, our method produces the lowest CP and the highest DVI, thus resulting in the lowest Ratio, for all the number of clusters. Our method achieves the best performance with clustering the students into 3 groups. Furthermore, we compute the CDs in the three resulting clusters for the four methods, shown in Fig. 4(c). From these results, metaDL yields a bigger CD than other methods.

Fig. 5 visualizes the three resultant clusters from metaDL. In Fig. 5, we also show the following three information: (1) Students whose Math grades in entrance exam are greater than 0.9 are labeled by big points. Note that we normalize the total grade into 1. (2) Students who like Math and Physics are labeled by blue boxes. (3) Students whose have high/medium/low grades in C language are colored by red/green/blue. From Fig. 5, we arrive at that the students who like Math and Physics and are good at Math generally obtain high grades in C language program, while the students who do not like Math and Physics and are weak at Math generally lead to low grades in C language program. Note that the grade in C language that is greater than 0.8 is defined as a high grade, while the grade that is less than 0.55 is defined as a low grade.

4.4. Student response prediction

In the applications of student knowledge diagnosis, an important practical problem is to predict whether a student could correctly response to an unseen question. Based on the predicted results, those questions that a student may make an incorrect response are recommended to help students improve their academic performance. We here determine a student's response on two factors: the knowledge points mastered by the student and the knowledge points involved in the objective question.

As a result, we design classification experiments on the three datasets and show the main steps as follows. We reformulate our data as a set of triplets, i.e., (studentID, questionID, response). There are $N \times M$ triplets. In the first step, we random sample a series of sub-datasets from the given dataset of different sizes. We in the second step employ the k-fold cross validation for each subdataset. In each fold, we train the model on training set and then predict each response of a triplet in test set. We finally obtain k prediction errors. What we do in the third step is computing the average error as the test error, where

$$\text{Prediction error} = \frac{\#\{\text{incorrect triplets in test set}\}}{\#\{\text{all triplets in test set}\}} \tag{25}$$

Besides, we run 20 times and achieve the average values to obtain a stable evaluation.

To determine the objective student's knowledge state, we calculate Euclidean distances between all students in the training set and the objective student based on the known responses, and then take the nearest neighbor knowledge state for the objective student. Since question's knowledge is also obtained from the training step, we here predict response by the simply dot multiplication of student's knowledge vector and question's knowledge vector. Besides, we employ the two-sample t-test and show all p-values for the statically significance.

4.4.1. FrcSub

On this data, we test different sub-datasets of different sizes in {50, 100, 250, 450, 535}. On each sub-dataset, we perform a five-fold cross-validation. That is, we employ the four algorithms per sub-dataset to obtain the student's prediction results and calculate the prediction error.

Fig. 6(a) shows the prediction errors of the four models on all sub-datasets. As is shown, metaDL results in the lowest error with a small standard derivation among the four methods. Given 50 samples, metaDL delivery 0.18 error that is about 0.04, 0.13 and

Table 4
Evaluation results of student grouping on NPU-C with the four methods.

	$c = 2$			$c = 3$			$c = 4$			$c = 5$		
	CP	DVI	Ratio	CP	DVI	Ratio	CP	DVI	Ratio	CP	DVI	Ratio
Original	2.057	0.426	4.828	2.467	0.302	8.169	2.495	0.333	7.491	1.157	0.316	3.660
SPARFA(5 features)	14.216	0.119	119.465	5.226	0.119	43.914	8.726	0.132	66.108	5.333	0.135	39.500
BOBCS(500 features)	0.064	0.432	0.149	0.141	0.432	0.326	0.081	0.431	0.189	0.053	0.457	0.115
MetaDL(500 features)	**0.047**	**0.531**	**0.087**	**0.044**	**0.536**	**0.085**	**0.053**	**0.535**	**0.098**	**0.047**	**0.536**	**0.088**

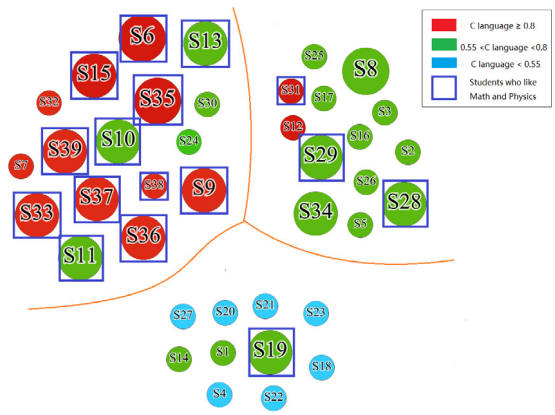


Fig. 5. Visualization of the student grouping results from metaDL on NPU-C. Points represent the students, while a big point has a good Math grade in the entrance exam. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
p-values between metaDL and other methods on FrcSub.

	metaDL *vs.* Original	metaDL *vs.* SPARFA	metaDL *vs.* BOBCS
50	4.409e−8	3.443e−17	1.875e−8
100	2.585e−7	8.765e−15	3.028e−8
250	1.846e−9	4.846e−16	1.901e−8
450	2.511e−9	1.809e−19	6.823e−9
535	7.481e−9	2.045e−18	2.619e−8

Table 6
p-values between metaDL and other Methods on ITI.

	metaDL *vs.* Original	metaDL *vs.* SPARFA	metaDL *vs.* BOBCS
50	9.475e−7	1.120e−6	2.089e−6
250	9.511e−8	1.975e−6	3.706e−7
350	1.978e−6	3.877e−7	2.243e−6
500	3.590e−7	9.955e−6	1.132e−6
1000	9.089e−7	2.201e−6	2.640e−6
1659	1.566e−6	8.596e−5	1.402e−5

0.29 less than SPARFA, Original and BOBCS respectively. Given 535 samples, metaDL model achieves the 0.168 error, indicating the effectiveness of our model. Besides, metaDL and SPARFA both have lower standard derivations than the method using original features.

Table 5 shows the *p*-values calculated from metaDL versus other methods using two-sample *t*-tests. From these results, all *p*-values are less than 0.001, which manifest these comparisons are statically significant.

4.4.2. ITI

On this dataset, we test different sub-datasets of different sizes in {50, 100, 350, 500, 1000, 1659}. The evaluations are computed by five-fold cross validations.

Fig. 6(b) describes the prediction errors for the four methods on different sub-datasets. On the sub-dataset of size 50, metaDL is about 0.14, 0.14 and 0.40 respectively less than Original, SPARFA and BOBCS in terms of prediction error. The *p*-values in Table 6 shows metaDL is significantly better than the compared methods. Besides, our method has the smallest standard deviation among the four methods. MetaDL and SPARFA have the same performance on the datasets with all samples, but metaDL is consistently better than SPARFA on all sub-datasets with small sample sizes.

4.4.3. NPU-C

Due to the small sample size in this dataset, we adopt the leave one out (LOO) method to evaluate the results. We test on the six sub-datasets of sizes in {5, 10, 15, 25, 35, 39}. Fig. 6(c) shows the prediction errors from different models on different sub-datasets. As is shown, metaDL gives rise to the lowest error on all datasets among the four methods. Specifically, metaDL achieves 26% error that is 10.4%, 10.4% and 24.7% less than Original, SPARFA and BOBCS respectively when the data size is 5. Besides, the *p*-values in Table 7 shows metaDL is significantly better in comparison with the other three methods. In real-world application, the given dataset is usually comprised of a small number of samples where our method could deliver a decent performance.

4.5. Student knowledge diagnosis results

This section analyzes the SKD results of metaDL, SPARFA and BOBCS on FrcSub, and then explains the results of the student

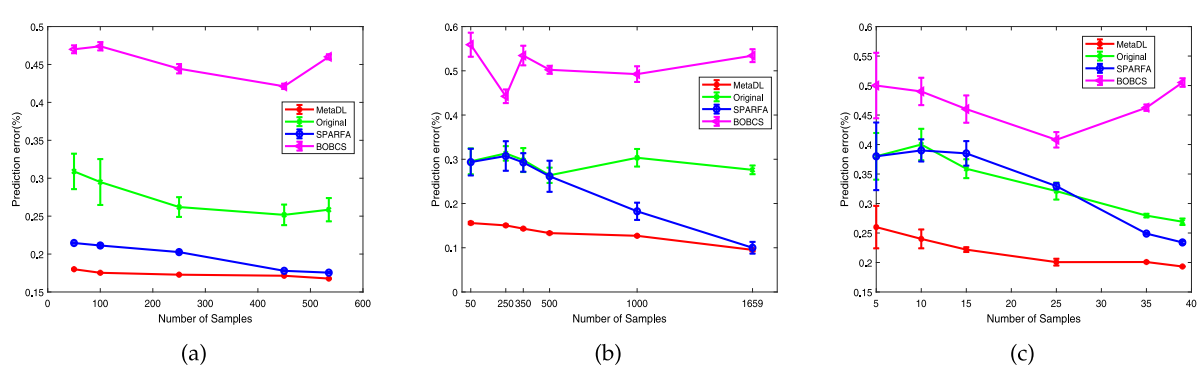


Fig. 6. The response prediction Results on (a) FrcSub, (b) ITI and (c) NPU-C with the four mentioned methods.

Table 7
p-values between metaDL and other methods on NPU-C.

	metaDL *vs.* Original	metaDL *vs.* SPARFA	metaDL *vs.* BOBCS
5	1.02e−5	1.575e−6	2.030e−5
10	7.211e−9	1.443e−9	9.769e−9
15	1.481e−9	6.102e−9	2.175e−8
25	4.476e−9	2.449e−10	1.076e−8
35	1.343e−13	1.115e−15	3.451e−9
39	1.736e−12	2.512e−15	2.121e−8

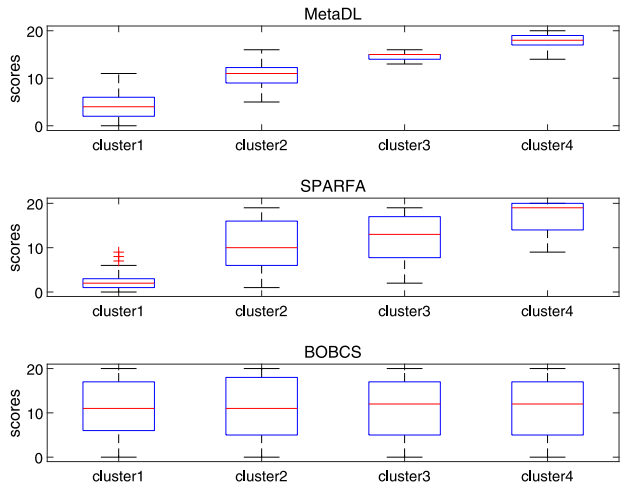


Fig. 7. The boxplots of student grouping results on FrcSub with MetaDL, SPARFA and BOCBCS.

knowledge diagnosis on NPU-C using expert concepts according to the formula in Section 3.4.

At first, the students in FrcSub are grouped into four clusters using the representations of the diagnosis results. Fig. 7 shows the boxplots of the scores of the students in each cluster. As is shown, metaDL delivers clear divisions of scores shown in Fig. 7 (MeatDL), while SPARFA and BOBCS lead to much overlaps between the four clusters shown in Fig. 7 (SPARFA and BOBCS). Thus, metaDL could result in a good explanation in education.

Second, on NPU-C dataset, we analyze the results on 25 expert concepts and show the top 3 concepts that are good or weak by mastery degrees in Tables 8 and 9. As is shown in Table 8, the top 3 mastery concepts of the three students are different on expert concepts, such as "comparison operators", "integer division" and "logic operation". While, Table 9 shows the top 3 non-mastery concepts of three students, such as "C language history", "data type", and "operators". Furthermore, Table 10 shows the weights of 3 concepts which most of the students are good at and weak at. The results suggest to supplementing the learnings on "array definition", "array initialization" and "logic operation".

Fig. 8(a) shows the relationship network between expert concepts labeled by pink and meta-knowledges from metaDL. From the relationship network, metaDL discovers the specific meta-knowledges labeled by yellow and the common knowledges labeled by green. That is, the expert concepts are composed of many the proposed meta-knowledges. As a result, metaDL deliveries the fine-grained representations of knowledge points.

Fig. 8(b) describes the diagnosis results on the proposed meta-knowledges, where the resultant 3 clusters are colored by deep green, purple, and pink. In Fig. 8(b), the yellow points are the specific meta-knowledges while the cyan points represent the common meta-knowledge. Wherein, the common meta-knowledges are shown in three cyan points, where most of them in the outer points are only shared with the students in a same cluster. As is result, the meta-knowledges can be categorized by the relationships: (1) the student-specific meta-knowledge that is only linked to a student; (2) the cluster-specific meta-knowledge that is shared in the same cluster; (3) the common meta-knowledge that is linked to students from different clusters.

Fig. 9 shows the relationships between the expert concepts and the students, where top 3 mastery concepts are linked to the corresponding students with weights in gray level. As can be seen, the resultant 3 clusters of students have their common expert concepts, while there are eight weak mastery concepts. It shows that metaDL could provide the mastery degree of the expert concept to all students. In addition, Fig. 10 depicts a student's mastery status of the 25 expert concepts. This student well masters the knowledge of "Sequence", "Pointer definition" and "Pointer operation".

5. Discussion and conclusion

Student knowledge diagnosis is a key problem of personalized learning to help teachers and students improve their education outcomes. The manual concept definition is a limitation

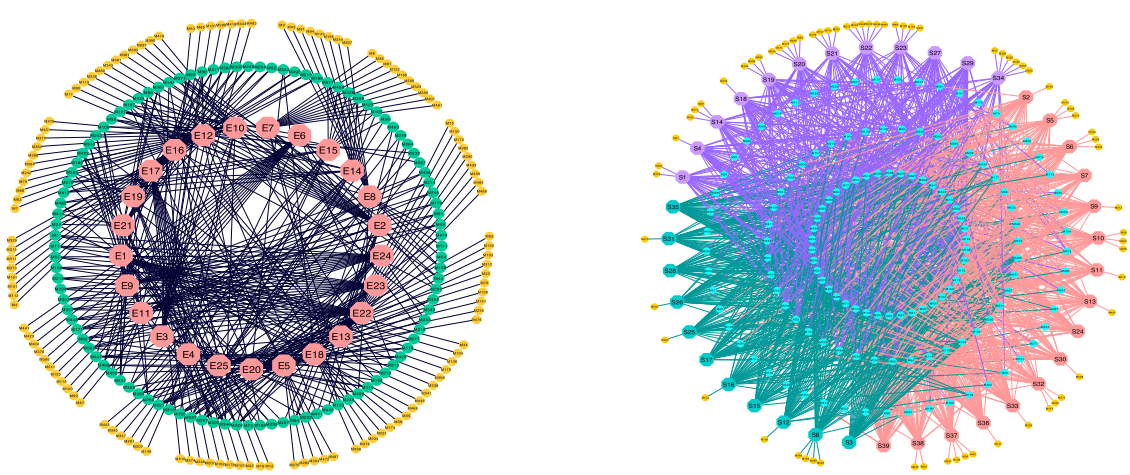


Fig. 8. Visualization results of metaDL method. (a) interprets the expert concepts in Q-matrix with meta-knowledge, where pink octagons are expert concepts and the rest points are meta-knowledges. (b) interprets the relationship between students and meta-knowledges, where points with label "S" represents students and others are meta-knowledges. The thicker the line represents the stronger the relationship between meta-knowledges and concepts or students. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

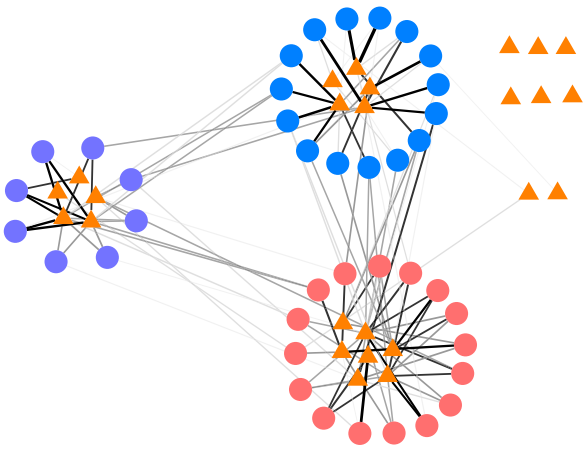


Fig. 9. Student-expert concepts association graph. Triangles represent expert concepts and points represent students, where three colors represent three clusters. The thicker the connection, the better the students master the concepts. Here, we only show the top 3 knowledge points mastered by each student.

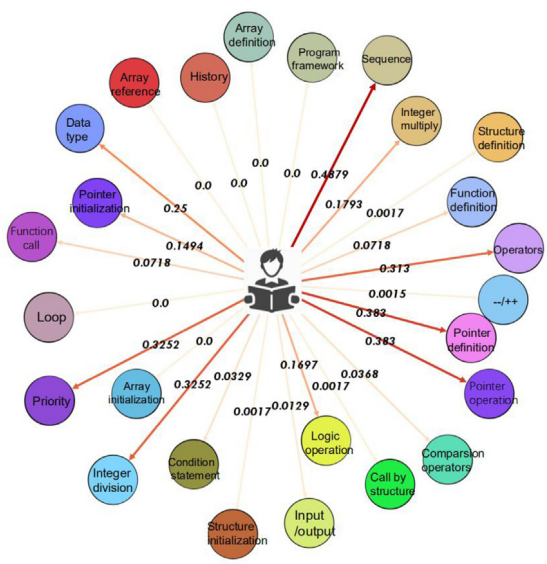


Fig. 10. A student's diagnosis results with metaDL on NPU-C. Circles indicate expert concepts, and the weights on lines correspond to mastery degrees.

Table 8
The top-3 well mastered expert-concepts of the three students.

student1	comparison operators	++/−−	condition statement
Weights	0.525	0.413	0.361
student2	integer division	comparison operators	structure definition
Weights	0.381	0.328	0.265
student3	logic operation	program framework	array index
Weights	0.531	0.394	0.342

Table 9
The top-3 weak mastered expert-concepts of the three students.

student1	C language history	data type	operators
Weights	0.001	0	0
student2	C language history	array definition	array initialization
Weights	0.001	0.001	0
student3	data type	condition statement	array definition
Weights	0	0	0

Table 10
The top-3 good mastered expert-concepts and the top-3 weak mastered expert-concepts in our class.

Good	Calculation priority	Function definition	Function call
Weights	0.411	0.538	0.562
Weak	array definition	array initialization	logic operation
Weights	0.589	0.535	0.333

to accurate explanation and student distinguishing in practical applications. Most recent studies are focused on learning complex and comprehensive concepts to replace expert concepts, while fail to effectively group students and predict new students' responses.

In this paper, we propose a data-driven method for student knowledge diagnosis in traditional course using the popular sparse learning theory. This study introduces a novel concept of meta-knowledge, which assumes knowledge can be divided into smaller meta-knowledges and uses the dictionary learning to learn the meta-knowledges from student's responses to exam papers. The proposed model is dubbed meta-knowledge dictionary learning, metaDL for short. To convince our proposed method, MetaDL was tested on two public datasets and a private dataset in comparison with other three related methods.

From the experiment results, metaDL could find the fine-gain knowledges, including student-specific meta-knowledges, cluster-specific meta-knowledges and common meta-knowledges, to have sparse student representations instead of traditional dense representations. The student-specific meta-knowledges might indicate the talent of a student, while the common meta-knowledge might indicate the core knowledge of a course. The cluster-specific meta-knowledges provide the suggestions for further learning. Based on this advantage, metaDL achieves significantly better results on student grouping and response prediction in comparison with other methods.

Specifically, metaDL results in the smallest CP, the largest DVI and thus the smallest Ratio values on the used datasets among all mentioned methods, manifesting that the meta-knowledge dictionary provides a good space to recognize the students with similar knowledge structure and separate the students with different knowledge structure. On the other hand, metaDL results in the lowest prediction errors on all experiment cases among all mentioned methods, showing that the knowledge representation from the diagnosis over the meta-knowledge dictionary has more robust and discriminative. On our private dataset, metaDL achieves Ratio $= 0.098$ that is about 7.4, 66, and 0.1 less than Original, BOBCS, and SPARFA for student grouping, while metaDL has 18% prediction error that is 0.06, 0.1 and 0.35 percents less than Original, SPARFA and BOBCS for response prediction. Besides, on the dataset of 5 samples, metaDL results in 26% prediction error that is 10.4, 10.4 and 24.7 percents less than Original, SPARFA and BOBCS for response prediction respectively.

However, there are two limitations in our studies and method as follows. (1) MetaDL fails to consider the data-missing case that would happen in real-world exams. The missing data might be tackled as incorrect responses and modeled by the noise matrix in metaDL. (2) The available dataset that contains expert concepts is small in surveys. MetaDL achieves better performance than other methods in the data-scares case in our experiment results. We leave the two limitations in our future studies. Besides, extracting the concepts from question texts is also our future work.

Overall, this study provides a new route to learn the fine-grain meta-knowledges for student knowledge diagnosis. This technique could facilitate the progress on personalized learning to enhance education outcomes and reduce education costs.

CRediT authorship contribution statement

Yupei Zhang: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing, Funding acquisition. **Huan Dai:** Methodology, Formal analysis, Investigation, Software, Data curation, Writing - original draft, Writing - review & editing. **Yue Yun:** Software, Validation, Data curation, Investigation. **Shuhui Liu:** Visualization, Writing - review & editing. **Andrew Lan:** Resources, Writing - review & editing. **Xuequn Shang:** Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is funded by the National Natural Science Foundation of China (Grants No. 61802313, U1811262, 61772426,) and the Fundamental Research Funds for Central Universities (Grant No. G2018KY0301).

References

[1] Chih-Ming Chen, Hahn-Ming Lee, Ya-Hui Chen, Personalized e-learning system using item response theory, Computers & Education 44 (3) (2005) 237–255.
[2] Karenne Hills, Kirsty Andersen, Samuel Davidson, Personalized learning and teaching approaches to meet diverse needs: a prototype tertiary education program, in: Reimagining Christian Education, Springer, 2018, pp. 233–257.
[3] Runze Wu, Qi Liu, Yuping Liu, Enhong Chen, Yu Su, Zhigang Chen, Guoping Hu, Cognitive modelling for predicting examinee performance, in: Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015.
[4] Trevor Bond, Zi Yan, Moritz Heene, Applying the Rasch Model: Fundamental Measurement in the Human Sciences, Psychology Press, 2013.
[5] Benjamin D Wright, Solving measurement problems with the rasch model, Journal of Educational Measurement 14 (2) (1977) 97–116.
[6] Erling B Andersen, A goodness of fit test for the rasch model, Psychometrika 38 (1) (1973) 123–140.
[7] Antonio Preti, Marcello Vellante, Donatella R Petretto, The psychometric properties of the "reading the mind in the eyes" test: an item response theory (irt) analysis, Cognitive Neuropsychiatry 22 (3) (2017) 233–253.
[8] R Chris Fraley, Niels G Waller, Kelly A Brennan, An item response theory analysis of self-report measures of adult attachment., Journal of personality and social psychology 78 (2) (2000) 350.
[9] Kazuhisa Noguchi, Eisuke Ito, Holonomic function of 2 parameter logistic model item response theory parameter estimation, in: Proceedings of the 10th International Conference on Education Technology and Computers, 2018, pp. 379–382.
[10] Youn-Jeng Choi, Natalia Alexeev, Allan S Cohen, Differential item functioning analysis using a mixture 3-parameter logistic model with a covariate on the timss 2007 mathematics test, International Journal of Testing 15 (3) (2015) 239–253.
[11] Stefan Hoffmann, Katja Soyez, A cognitive model to predict domain-specific consumer innovativeness, Journal of Business Research 63 (7) (2010) 778–785.
[12] Peida Zhan, Hong Jiao, Manqian Liao, Yufang Bian, Bayesian dina modeling incorporating within-item characteristic dependency, Applied psychological measurement 43 (2) (2019) 143–158.
[13] Jimmy De La Torre, Dina model and parameter estimation: a didactic, Journal of educational and behavioral statistics 34 (1) (2009) 115–130.
[14] Ping Chen, Tao Xin, Chun Wang, Hua-Hua Chang, Online calibration methods for the dina model with independent attributes in cd-cat, Psychometrika 77 (2) (2012) 201–222.
[15] Huilin Chen, Jinsong Chen, Exploring reading comprehension skill relationships through the g-dina model, Educational Psychology 36 (6) (2016) 1049–1064.
[16] Bor-Chen Kuo, Chun-Hua Chen, Jimmy de la Torre, A cognitive diagnosis model for identifying coexisting skills and misconceptions, Applied Psychological Measurement 42 (3) (2018) 179–191.
[17] Dimiter M Dimitrov, Dimitar Atanasov, Group comparisons on cognitive attributes using the least squares distance model of cognitive diagnosis, Pliska Studia Mathematica Bulgarica 22 (1) (2013) 33p–40p.
[18] Jingchen Liu, Gongjun Xu, Zhiliang Ying, Data-driven learning of q-matrix, Applied psychological measurement 36 (7) (2012) 548–564.
[19] Chia-Yi Chiu, Statistical refinement of the q-matrix in cognitive diagnosis, Applied Psychological Measurement 37 (8) (2013) 598–618.
[20] Hans-Friedrich Köhn, Chia-Yi Chiu, How to build a complete q-matrix for a cognitively diagnostic test, Journal of Classification 35 (2) (2018) 273–299.
[21] Andrew S Lan, Andrew E Waters, Christoph Studer, Richard G Baraniuk, Sparse factor analysis for learning and content analytics, The Journal of Machine Learning Research 15 (1) (2014) 1959–2008.
[22] Zameer Gulzar, A Anny Leema, Gerard Deepak, Pcrs: personalized course recommender system based on hybrid approach, Procedia Computer Science 125 (2018) 518–524.
[23] Frank B Baker, The basics of item response theory, ERIC, 2001.
[24] I Toić, P Frossard, Dictionary learning: what is the right representation for my signal, IEEE Signal Processing Magazine. v28 i2 (2011) 27–38.
[25] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, Online dictionary learning for sparse coding, in: Proceedings of the 26th annual international conference on machine learning, 2009, pp. 689–696.
[26] Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, Francis R Bach, Supervised dictionary learning, in: Advances in neural information processing systems, 2009, pp. 1033–1040.
[27] Rikkert M van der Lans, Wim JCM van de Grift, K Van Veen, Developing an instrument for teacher feedback: using the rasch model to explore teachers' development of effective teaching strategies and behaviors, The journal of experimental education 86 (2) (2018) 247–264.
[28] Song Cheng, Qi Liu, Enhong Chen, Zai Huang, Zhenya Huang, Yiying Chen, Haiping Ma, Guoping Hu, Dirt: deep learning enhanced item response theory for cognitive diagnosis, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 2397–2400.
[29] Abd-Krim Seghouane, Asif Iqbal, Karim Abed-Meraim, Sequential structured dictionary learning for block sparse representations, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 3397–3401.
[30] Selen Ayas, Murat Ekinci, Single image super resolution using dictionary learning and sparse coding with multi-scale and multi-directional gabor feature representation, Information Sciences 512 (2020) 1264–1278.
[31] Yupei Zhang, Ming Xiang, Bo Yang, Graph regularized nonnegative sparse coding using incoherent dictionary for approximate nearest neighbor search, Pattern Recognition 70 (2017) 75–88.
[32] Xuefeng Chen, Zhaohui Du, Jimeng Li, Xiang Li, Han Zhang, Compressed sensing based on dictionary learning for extracting impulse components, Signal Processing 96 (2014) 94–109.
[33] H Zayyani, M Korki, F Marvasti, Bayesian hypothesis testing for one bit compressed sensing with sensing matrix perturbation, Scientia Iranica 25 (6) (2018) 3628–3633.
[34] Yael Yankelevsky, Michael Elad, Dictionary learning for high dimensional graph signals, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 4669–4673.
[35] HaoXuan Ni, Jinzuo Ye, Dehui Xiang, Yang Du, Xinjian Chen, Xiang Deihui, Jie Tian, A fast reconstruction algorithm for fluorescence molecular tomography via multipath subspace pursuit method, in: Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging, vol. 10578, International Society for Optics and Photonics, 2018, p. 1057810.
[36] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, Sebastian Thrun, Dermatologist-level classification of skin cancer with deep neural networks, nature 542 (7639) (2017) 115–118.
[37] Hadi Zayyani, Mehdi Korki, Farrokh Marvasti, Dictionary learning for blind one bit compressed sensing, IEEE Signal Processing Letters 23 (2) (2015) 187–191.
[38] Jeffrey D Blanchard, Jared Tanner, Ke Wei, Cgiht: conjugate gradient iterative hard thresholding for compressed sensing and matrix completion, Information and Inference: A Journal of the IMA 4 (4) (2015) 289–327.
[39] Karin Schnass, Convergence radius and sample complexity of itkm algorithms for dictionary learning, Applied and Computational Harmonic Analysis 45 (1) (2018) 22–58.
[40] Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, Yike Guo, et al., Dagan: deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction, IEEE transactions on medical imaging 37 (6) (2017) 1310–1321.
[41] Yu Wang, Wotao Yin, Jinshan Zeng, Global convergence of admm in nonconvex nonsmooth optimization, Journal of Scientific Computing 78 (1) (2019) 29–63.
[42] Jian-Feng Cai, Emmanuel J Candès, Zuowei Shen, A singular value thresholding algorithm for matrix completion, SIAM Journal on optimization 20 (4) (2010) 1956–1982.

[43] Isidora Stankovic, Miloš Brajovic, Miloš Dakovic, Ljubiša Stankovic, Complex-valued binary compressive sensing, in: 2018 26th Telecommunications Forum (TELFOR), IEEE, 2018, pp. 1–4.
[44] Shengzheng Wang, Baoxian Ji, Jiansen Zhao, Wei Liu, Tie Xu, Predicting ship fuel consumption based on lasso regression, Transportation Research Part D: Transport and Environment 65 (2018) 817–824.
[45] M Alibuhtto, N Mahat, Distance based k-means clustering algorithm for determining number of clusters for high dimensional data, Decision Science Letters 9 (1) (2020) 51–58.
[46] Chen Wu, Chao Yang, Shenglan Ma, Xiaoliang Xu, Feasibility study on grouting compactness detection in sleeves using piezoelectric transducers, Applied Sciences 10 (1) (2020) 149.