Refinement of a Q-matrix with an ensemble technique based on multi-label classification algorithms

Sein Minn[1], Michel C. Desmarais[1], and ShunKai Fu[2]

[1] Polytechnique Montreal
{sein.minn,michel.desmarais}@polymtl.ca
[2] Huaqiao University
fusk@hqu.edu.cn

Abstract Many algorithms and tools exist to help an expert map exercises and tasks to underlying skills. The last decade has witnessed a wealth of data driven approaches aiming to refine expert-defined mappings of tasks to skill. This refinement can be seen as a classification problem: for each possible mapping of task to skill, the classifier has to decide whether the expert's advice is correct, or incorrect. Whereas most algorithms are working at the level of individual mappings, we introduce an approach based on a multi-label classification algorithm that is trained on the mapping of a task to all skills simultaneously. The approach is shown to outperform the existing Q-matrix refinement techniques (such as MinRSS, MaxDiff, Matrix Factorization).

Keywords: Student model, Skills modeling, Psychometrics, Q-matrix validation, Multi-label skills assessment

1 Introduction

Intelligent tutoring systems rely on efficient methods to assess the skills to perform tasks. These skills can involve factual knowledge, deep understanding of abstract concepts, general problem solving abilities, practice at recognizing patterns and situations, etc. Furthermore, a designer of a learning environment may focus on a particular subset of these skills. It might be the subset that is deemed appropriate for 10–12 years old kids endowed with a specific training. Or it might be a subset that more closely relates to a given topical or pedagogical perspespective at the expense of alternative perspectives. For example a tutor may not care much about general problem solving abilities that require months to acquire, and focus on factual knowledge and rules that are easier to teach and assess, even though both the problem solving skills and factual knowledge are involved in the training and assessment material.

Whathever the motivation is for defining the skills behind the successful completion of tasks, a first point to emphasize is that for the same tasks, one skill definition may be considered appropriate for one context whereas another will be

required for another context. A second point to emphasize is that the definition of skills behind tasks, or the converse, the definition of tasks for a given set of skills, are non trivial and error prone processes.

Therefore, tools to help a tutor, or a designer of a learning environment, validate a given mapping of skills to tasks would be highly valuable. Let us refer to this endeavor as the problem of *Q-matrix refinement*, where the Q-matrix represents the mapping of tasks to skills.

In this paper, we present a framework to help validate a Q-matrix called Multi-Label Skills Refinement (MLSR). We describe the method, setup, analysis, and results of a performance assessment of Q-matrix refinement.

This approach can be considered an ensemble technique, since it combines refinements obtained from different algorithms to calculate its own refinements: Minimal Residual Sum Square (MinRSS), Maximum Difference (MaxDiff) and Conjunctive Alternating Least Square Factorization (ALSC). In addition, the approach uses features obtained from a large number of simulations with the refinements algorithms, and in particular an indicator of each algorithm's error rate over a given cell of the Q-matrix. The error rate computed from these simulations by using synthetic data, for which the ground thruth is known.

The rest of this paper is organized as follow. Section 2 reviews the related work on the Q-matrices and techniques to validate them from data. Section 3 combining techniques with multi-label classification, Section 4 presents the measurement methods. Experimental results are found in Section 5 and Section 6 concludes and discusses future prospective.

2 Q-Matrices and Related work

Modelling and predicting how human beings learn is dealing with many fields as diverse as neuroscience, education, psychology, and cognitive science. So assessing latent skills influenced by complex macro level interactions is non obvious and non traceable. The challenges present are further exposed on the micro level interactions and factors. These all latent skills extracted from response matrix (performance of students: students x tasks) and can decompose into two important matrices: Q-matrix (tasks x skills) and P-matrix (students x skills) [1,2,14]. These two matrices contains all information about how students response and how much they will be able to perform their tasks.

Q-matrix : A mapping of items to skills is termed a Q-matrix [10,5,7]. An example of a 11 items and 5 skills Q-matrix is given below, where item 4 requires skill 1 only, whereas item 11 requires skill 2 and 4. If all specified skills are required to succeed the item, the Q-matrix is conjunctive. If a any of the required skill is sufficient to the item success is disjunctive.The compensatory corresponds to the case where each required item increases the chances of success in some way [7].

	s_1	s_2	s_3	s_4	s_5
i_1	1	1	0	1	0
i_2	0	1	1	1	0
i_3	1	0	1	0	0
i_4	1	0	0	0	0
i_5	0	0	1	0	0
i_6	0	1	0	1	1
i_7	1	1	1	1	0
i_8	0	1	1	1	1
i_9	0	1	1	1	0
i_{10}	0	0	1	1	0
i_{11}	0	1	0	1	0

Table 1. Example of Q-matrix

2.1 Q-matrix Refinement techniques

Whereas we find a number of techniques to derive Q-matrices entirely from from data (for eg. [1,2,14]), the current study focuses on a related problem: refining expert-given Q-matrices from data. The two techniques are closely related. The main difference can generally be considered as one of starting points: entirely data-driven Q-matrix definition starts from a random state, or from some predetermined state, whereas refinement techniques start from the expert's Q-matrix. However, very often, the general algorithms are the same.

We chose to use three Q-matrix refinement techniques that were studied in [7,4] for the purpose of comparison. They are described below.

MinRSS : Minimal Residual Sum Square (MinRSS) is from [3,7].A given Q-matrix, It provides an ideal response pattern for a given a student skills mastery profile. This ideal response pattern sigificantly relies on Q-matrix given student profile. That is, if there are no slip and guess factors, then the response pattern for every profile of student is fixed. Measuring the difference between the real response pattern and the ideal response pattern give us a value to fit for the Q-matrix. The most common meausrement for vector is Hamming distance, that is

$$d_h(r,\eta) = \sum_{j=1}^{J} |r_j - \eta_j| \tag{1}$$

where r is the real response vector while η is the ideal response vector. J is the number of latent skills. chiu et al. [3] leads us to a more refined metric. The idea is if an item has a smaller entropy, then it should be given higher weight. The formula is

$$d_{\omega h}(r,\eta) = \sum_{j=1}^{J} \frac{1}{\bar{p_j}(1-\bar{p_j})} |r_j - \eta_j| \tag{2}$$

where $\bar{p_j}$ is the proportion of correct answers of item j. Equipped with this metric, we can find the most approximate ideal response matrix and then find the corresponding profile matrix A. First, a squared sum of errors for each item

k can be computed by

$$RSS_k = \sum_{i=1}^{N} (r_{ik} - \eta_{ik})^2 \tag{3}$$

where N is the number of examinees or respondents. Then, the item with the highest RSS is chosen to minimize with it's correspondent q-vector. All the other possible q-vectors are calculated their RSS and the q-vector giving the lowest RSS is chosen to replace the original one. This method called minRSS. The Q-matrix is updated during the whole process repeated, but the previously changed q-vector is eliminated from the next round of running. The whole procedure terminates until the RSS for each item no longer changes. This method was proposed by [17] to yield good performance under different underlying conjunctive models.

MaxDiff : According to DINA model, for every item j, there are two model parameters called, slip s_j and guess g_j. de la Torre et al. [14,7] proposed that a correctly specified q-vector for item j should maximize the difference of probabilities of correct response between examinees who possess all the required skills and those who do not. That is, q_j is the correct q-vector if

$$\begin{aligned} q_j &= \arg\max_{\alpha_l} [P(X_j = 1|\xi_{ll'} = 1) - P(X_j = 1|\xi_{ll'} = 0)] \\ &= \arg\max_{\alpha_l} [\delta_{jl}] \end{aligned} \tag{4}$$

where $\xi_{ll'} = \prod_{k=1}^{K} \alpha_{l'k}^{\alpha_{lk}}$ for K total number of skills. An interesting observation is that since $P(X_j = 1|\xi_{ll'} = 1) = 1 - s_j$ and $P(X_j = 1|\xi_{ll'} = 0) = g_j$, then

$$q_j = \arg\max_{\alpha_l} [1 - (s_j + g_j)]$$

that is, repeatedly maximizing the difference is equivalent to minimize the sum of the slip and guess parameters iteratively. A original idea is to test all q-vectors to find the maximum δ_{jl} but that is computationally unefficient. de la Torre et al. [14] proposed a greedy algorithm that adds skills into a q-vector sequentially. First, δ_{jl} is calculated for all q-vectors which contains only one skill and the one with biggest δ_{jl} is chosen. Then, δ_{jl} is calculated for all q-vectors which contains two skills including the previously chosen one. Again the q-vector with the biggest δ_{jl} is chosen. This whole process is repeated until no addition of skills increases δ_{jl}. However, this algorithm requires knowledge of s_j and g_j in advance. They are calculated by EM (Expectation Maximization) algorithm [15].

ALSC : ALSC (Conjunctive Alternating Least Square Factorization) is a common matrix Factorization (MF). Desmarais et al. [6,7] proposed to factorize student test results into a Q-matrix and a profile matrix by using ALSC.

Contrary to the other two methods, it has no slip and guess parameters. ALSC decomposes the results matrix $\mathbf{R}_{m\times n}$ of m items by n students as the inner product two smaller matrices:

$$\neg\mathbf{R} = \mathbf{Q}\,\neg\mathbf{S} \tag{5}$$

where $\neg\mathbf{R}$ is the negation of the results matrix (m items by n students), $\mathbf{Q}$ is the m items by k skills Q-matrix, and $\neg\mathbf{S}$ is negation of the the mastery matrix of k skills by n students (normalized for rows columns to sum to 1). By negation, we mean the 0-values are transformed to 1, and non-0-values to 0. Negation is necessary for a conjunctive Q-matrix.

The factorization consists of alternating between estimates of $\mathbf{S}$ and $\mathbf{Q}$ until convergence. Starting with the initial expert defined Q-matrix, $\mathbf{Q}_0$, a least-squares estimate of $\mathbf{S}$ is obtained:

$$\neg\hat{\mathbf{S}}_0 = (\mathbf{Q}_0^{\mathrm{T}}\,\mathbf{Q}_0)^{-1}\,\mathbf{Q}_0^{\mathrm{T}}\,\neg\mathbf{R} \tag{6}$$

Then, a new estimate of the Q-matrix, $\hat{\mathbf{Q}}_1$, is again obtained by the least-squares estimate:

$$\hat{\mathbf{Q}}_1 = \neg\mathbf{R}\,\neg\hat{\mathbf{S}}_0^{\mathrm{T}}\,(\neg\hat{\mathbf{S}}_0\,\neg\hat{\mathbf{S}}_0^{\mathrm{T}})^{-1} \tag{7}$$

iteratively until convergence. Alternating between equations (6) and (7) yields progressive refinements of the matrices $\hat{\mathbf{Q}}_i$ and $\hat{\mathbf{S}}_i$ that more closely approximate $\mathbf{R}$ in equation (5). The final $\hat{\mathbf{Q}}_\mathbf{i}$ is rounded to yield a binary matrix.

3 Multi-Label Skills Refinement

Each of the three techniques described above, MinRSS, MaxDiff, and ALSC, uses a substantially different algorithm from the others to refine a Q-matrix. In that respect, their respective outcome may be complementary, and we can hypothesize that they can be combined to provide a more reliable output than any single one. Furthermore, some algorithms are more effective in general, but may not be the best performer in all context. Defining the features that that allows learning which algorithm provides the most reliable outcome in a given context is another objective of combining these techniques.

We first describe the data on which the multi-label skill refinement techniques are trained, and then describe the two algorithms that use this data.

3.1 Data to train the muti-label skills refinement algorithms

Let us introduce some notations used in this study. Given an instance X and its associated label set $l_i \subset |L|$, where its l_i component of $|L|$ takes the value of 1 if $l_i \in |L|$ and 0 otherwise. In addition, let $N(x)$ denote the set of x identified in the training set.

Table 2 contains an excerpt of data used to train the multi-lable skills refinenent algorithms. Each line is a record for a single item to skills mapping.

The rightmost columns contain the true labels. The left columns contain the suggested refinements from the different algorithms and factors that may provide information about the most reliable technique refinement in a given context. They are:

Stickiness is the proportion of more likely possible truly classified cells in iteratively refined matrices from any type of pertubation.

$$Stickiness = \frac{\sum_{t=1}^{T}(r \neq p)}{T} \tag{8}$$

where r is the cells in refined matrix and p is the cells in pertubated matrix. T is the number of times they iterate according to number of rounds we predefined. (In this experiment, we used total number of cells in matrices are defined as their round number).

Skills per row indicates the number of skills required for Items. An item may contain one or more skills.

Skills per column is the sum of the skills per columns. an indicator of how often this skill is required by the different items of the Q-matrix.

Item	Min.s1	Min.s2	Min.s3	St.Min.s1	St.Min.s2	St.Min.s3	...	**true.s1**	**true.s2**	**true.s2**
1	1	1	0	0.04	0.04	0.00	...	1	1	0
2	0	1	0	0.00	0.06	0.10	...	0	1	1
3	1	1	1	0.20	0.05	0.00	...	1	0	1
4	1	0	0	0.04	0.04	0.20	...	1	0	0
5	1	0	1	0.00	0.04	0.04	...	1	0	1

Table 2. Example of data set used for multi-label classification

3.2 Multi-Label Skills Algorithms

We transform the outputs and latent factors from three data driven techniques into multi-label classification problem. After that we use accumulated data from synthetic 1000 permutated matrices for training and use real data for testing, the procedures for data generation of training and testing are shown in Fig: 2. Finally we use multi-label classification for prediction of skills for each items. Generality of multi-label problems significantly makes it more complex to solve than traditional single-label (two-class or multi-class) problems. Only a few studies on multi-label learning are reported in the literature, which mainly concern the problems of text categorization, bioinformatics and scene classification. In this study, we conduct two multi-label classification methods: binary relevance method (Classifier chain method) [11] by using Naive Bayes classifier, and RAndom k-labELsets(Ensemble method) [16] by using J48 decision tree algorithm for our skills refinement tasks.

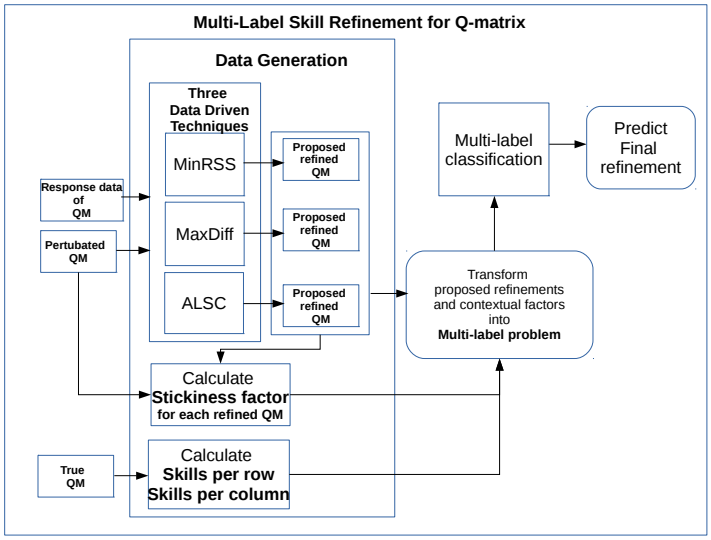

Figure 1. Refinement Procedure of each Q-Matrix QM_i

Binary Relevance method with Naive Bayes The strategy of problem transformation is to use the one-against-all strategy by converting the multi-label problem into several binary classification problems. This approach is known as the binary relevance method (BR) [11]. A method closely related to the BR method is the Classifier Chain method (CC) proposed by Read et al. [11]. This method involves Q binary classifiers linked along a chain. BR transforms any multi-label problem into one binary problem for each label. Hence this method trains $|L|$ binary classifiers $C_1, ..., C_{|L|}$. Each classifier C_j is responsible for predicting the $0/1$ association for each corresponding label $l_j \in L$.

BR with Naive Bayes (NB) method make NB classifiers are linked in a chain, such that classifier for l_i in chain considers the classes predicted $l_1, l_2, ..., l_{i-1}$ from the previous classifiers as additional attributes. Thus, the feature vector for each binary classifier is extended with the class values (labels) of all previous classifiers in the chain. Each classifier in the chain is trained to learn the association of label L_i given the features augmented with all previous class labels in the chain, $C_1; C_1; C_2; ...; C_{|L|}$. At classification time, the process starts at C_1, and propagates the predicted classes along the chain such that for C_i it computes:

$$P(l_i) = \arg\max_{l_i} P(l_i|X, l_1, l_2, ..., l_{i-1}) \tag{9}$$

RAndom k-labELsets with J48 The ensemble methods for multi-label learning are developed on top of the common problem transformation or algorithm

adaptation methods. The most well known problem transformation ensembles are the RAndom k-labELsets (RAkEL) system by Tsoumakas et al. [16]. RAkEL constructs each base classifier by considering a small random subset of labels and learning a single-label classifier for the prediction of each element in the power-set of this subset that transformed form multi-label problem.

In here, single-label J48 classifier, an optimized implementation of the C4.5 or improved version of the C4.5. J48 constructs Decision tree as an output. A Decision tree have same structure as tree that have different types of node, such as root node, intermediate nodes and leaf node. Each node in the tree containing constraints and that constraint leads to our result as name as decision tree. Decision tree divides the input space of a dataset into mutual exclusive areas, where each area having a class of labels, a value or an action to describe or elaborate its nodes of data. Splitting criterion is used in decision tree to determine which attributes are the optimal to split into portion of tree given training data.

4 Evaluation measurement principle

The evaluation of methods for multi-label data requires different measurement than those used in the case of single label data. For the definitions of these measures we will consider an evaluation data set of multi-label examples $(x_i, Y_i), i = 1...m$, where $Y_i \subseteq L$ is the set of true labels and Z_i is the set of predicted labels. This section presents two group of measurements [8] that will be used in this experiment for the evaluation of our method.

- Example based measurement: are calculated over all examples of the evaluation data set, that based on the average differences of the actual and the predicted sets of labels.

4.1 Example based Measurement

Hamming Loss : is measurement of how many times an instance label set is misclassified, i.e. a label not belonging to the instance is predicted or a label belonging to the instance is not predicted. The performance is perfect when $HammingLoss = 0$; the smaller the value of $HammingLoss$, the better the performance:

$$HammingLoss = \frac{1}{m} \sum_{i=1}^{m} \frac{|Z_i \Delta Y_i|}{M} \tag{10}$$

where Δ stands for the symmetric difference between two laebel sets. which is the theoretic equivalent of the exclusive disjunction (XOR operation) in Boolean logic for sets.

Subset Accuracy To calculate the accuracy of vector of labels is truely classified. $SubsetAccuracy$ is defined as follows:

$$SubsetAccuracy = \frac{1}{m}\sum_{i=1}^{m} I(Z_i = Y_i) \tag{11}$$

Example based F-score : are calculated based on the average differences of the actual and the predicted sets of labels over all examples of the evaluation data set. The performance is perfect when $ExamplebasedF - score = 1$; the bigger the value ,the better the performance:

$$OneError = \frac{1}{m}\sum_{i=1}^{m} \frac{2|Y_i \cap Z_i|}{|Z_i| + |Y_i|} \tag{12}$$

5 Experimental Study

For the sake of comparison, we use the same datasets as the ones used in Desmarais et al. (2015) [13,7]. Table 3 provides the basic information and source of each dataset.

Table 3. Q-matrix for validation & explantion of category

Q-matrices	**Number of**			**Description**
	Skills	Items	Cases	
QM1	3	11	536	Expert driven from [9]
QM2	5	11	536	Expert driven from [14]
QM3	3	11	536	Expert driven from [12]
QM4	3	11	536	Data driven, SVD based

and number of $1, 2, 3, 4$ along with algorithms in results of our experiment represent the categories of number of features they contained respectively.

1 : contains item number, outputs from three different basic algorithms
2 : contains item number, stickiness factors from three different algorithms
3 : contains combination of item number, outputs, row sum and column sums.
4 : contains combination of item number, outputs,stickiness factors, row sum and column sums.

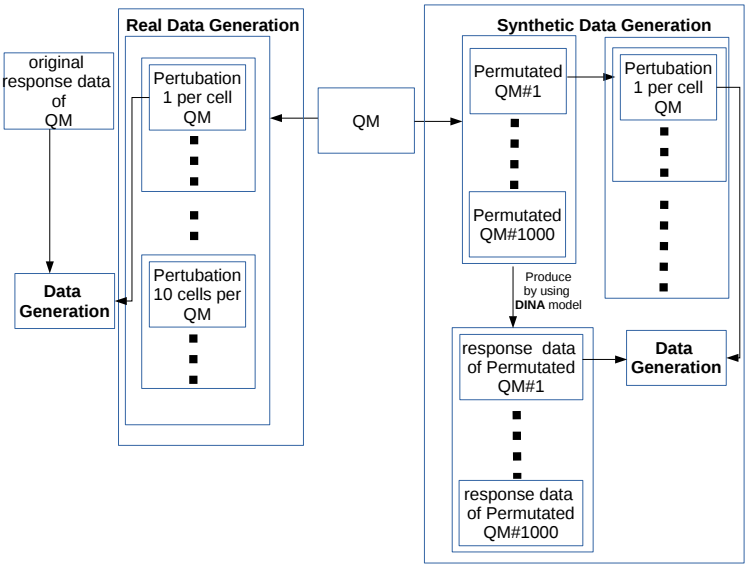

Figure 2. Data Generation Procedure of each Q-Matrix QM_i

For this experiment, We produced synthetic datasets that contains outputs and latent factors from 1000 permutated matrices of QM-1 (3 skills) and QM-2 (5 skills) by using 400 response data and each permutated matrix is refined from matrices with one cell pertubation for every single cell in it. We trained the classifier with synthetic datasets from one cell pertubated Q-matrices and test over real datasets of multiple pertubted cells (ranged from 1 to 10).

QM	MinRSS	MaxDiff	ALSC	RAkEL(1)	BR(1)	RAkEL(2)	BR(2)	RAkEL(3)	BR(3)	RAkEL(4)	BR(4)
qm1	0.53 ± 0.00	0.09 ± 0.00	0.54 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
qm2	0.42 ± 0.00	0.41 ± 0.00	0.44 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.09 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
qm3	0.63 ± 0.00	0.64 ± 0.00	0.55 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.03 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
qm4	0.58 ± 0.00	0.59 ± 0.00	0.53 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.25 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

Table 4. Hamming Loss result of Synthetic data

In this experiment, the result of various algorithms with different types of individual or combination with outputs and latent features are represented. For testing synthetic datasets, we use 10 fold cross validation and, we use train/test setting in testing our real datasets. We proved that optimal performance of our refinement methods with series of measurements for synthetic dataset and testing with real datasets. We use Hamming loss, Subset Accuracy and Example based F-measure to measure prediction performance on vector of skills.

QM	MinRSS	MaxDiff	ALSC	RAkEL(1)	BR(1)	RAkEL(2)	BR(2)	RAkEL(3)	BR(3)	RAkEL(4)	BR(4)
qm1	0.19 ± 0.00	0.85 ± 0.00	0.18 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	1.00 ± 0.00	0.99 ± 0.00
qm2	0.24 ± 0.00	0.25 ± 0.00	0.13 ± 0.00	1.00 ± 0.00	0.93 ± 0.00	1.00 ± 0.00	0.73 ± 0.00	1.00 ± 0.00	0.94 ± 0.00	1.00 ± 0.00	0.98 ± 0.00
qm3	0.00 ± 0.00	0.02 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.91 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
qm4	0.07 ± 0.00	0.08 ± 0.00	0.04 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	1.00 ± 0.00	0.50 ± 0.05	1.00 ± 0.00	0.98 ± 0.00	1.00 ± 0.00	0.98 ± 0.00

Table 5. SubSet Accuracy result of Synthetic data

QM	MinRSS	MaxDiff	ALSC	RAkEL(1)	BR(1)	RAkEL(2)	BR(2)	RAkEL(3)	BR(3)	RAkEL(4)	BR(4)
qm1	0.54 ± 0.00	0.90 ± 0.00	0.54 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
qm2	0.68 ± 0.00	0.71 ± 0.00	0.64 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	0.92 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
qm3	0.06 ± 0.00	0.10 ± 0.00	0.16 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.96 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
qm4	0.37 ± 0.00	0.37 ± 0.00	0.42 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.66 ± 0.02	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00

Table 6. Macro averaged F-measure result of Synthetic data

The experimental results on each evaluation criterion are reported in Tables [4,5,6] for synthetic data and in Figures [3,4,5] for real data, where the results from three basic algorithms are in black lines and multi-label skill refinement methods are in color lines on each Q-matrix. According to results, applying multi-label skill refinement outperforms than three basic algorithms. For Synthetic data, most of multi-label skill refinement methods can recover over 99% for all Q-matrices and even the performance reaches 100% of subset accuracy and macro averaged F-measure.

For real data, Among those multi-label skill refinement methods, BR by using category 1 and category 2 of data show the best performance. The performance of all these methods will decline with numbers of pertubation because of the synthetic data sets that we trained with one cell pertubated dataset can not handle the noises from real data sets with multiple cells (1 to 10) pertubation. One caveat of our method is, it can not be used under condition of the whole column is 1, For instance, skill-1 is required for all tasks. we do not need to consider whether this skill-1 is required for which tasks. In our experiment, Skill-1 of QM-1 is such type of skill, so we ignore that skill when we do prediction and evaluation. When we have few number of labels, the method didn't show its optimal performance. If we compare QM-2 to QM-1, QM-3 and QM-4. It show better performance in QM2 compared to others because of QM-2 have 5 skills and others only have 3 skills. In QM-1, we only predict 2 skills for multi-label classification, it shows lower performance than QM-3 and QM-4. So we can conclude that if we have to predict more skills, the methods are more effective.

6 Conclusion & Future Work

In this paper, we represent the multi-label skill refinement method, that combines three data driven techniques and each of two multi-label classification techniques to assess the skills required for tasks of students. Experiment with 3 expert driven Q-matrices and 1 Q-matrix driven form SVD, shows proposed refinement methods outperform than three well known algorithms in literature and can recover more accurate skill vector in Q-matrix when it only contains noises as same as when

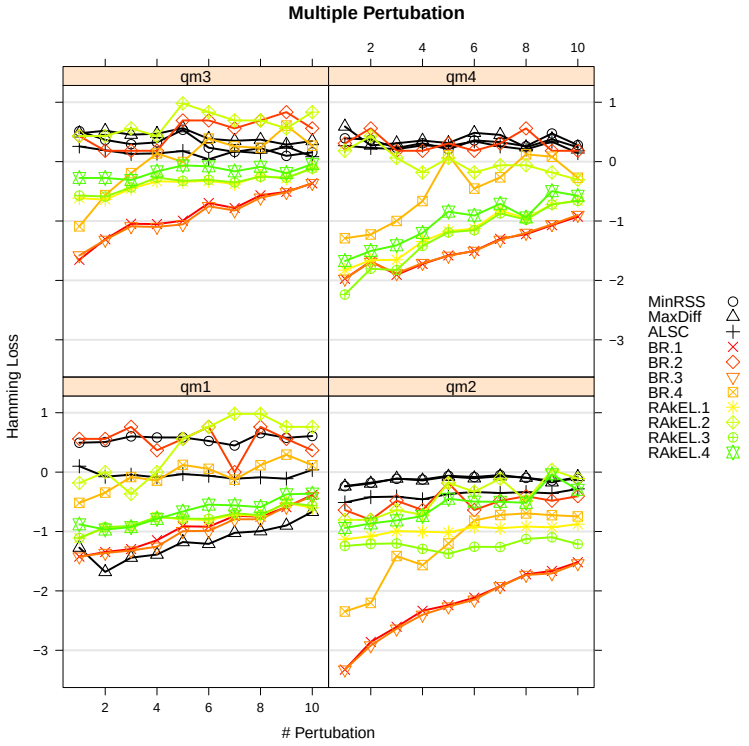

Figure 3. Logit value of Hammig loss of real data

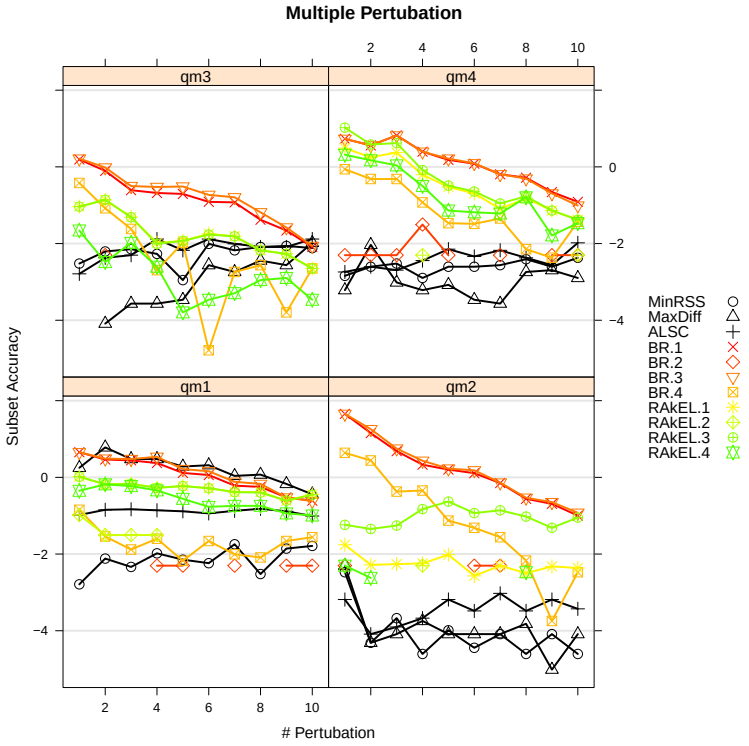

Figure 4. Logit value of Subset Accuracy of real data

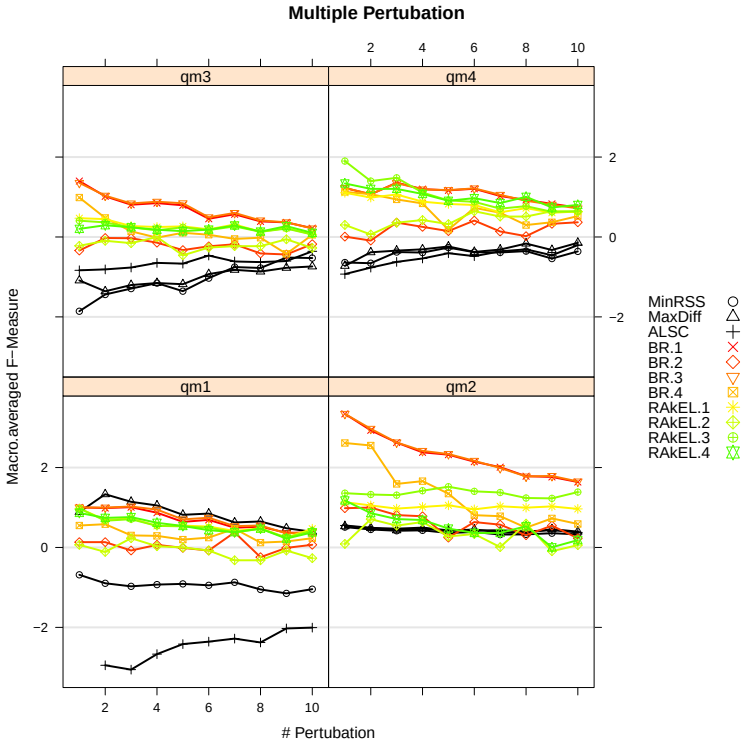

Figure 5. Logit value of Example-based F-measure of real data

we trained. Therefore, it is interesting to see how this method would perform in when we give enough noises in training, dealing with non-binary Q-matrices, skills varying over time.The another biggest challenge is finding out micro level influences over these skills. This particular study was highly facilitated by the CDM [12]and NPCD packages which provided both the code for three basic data driven techniques and the data, and mulan [8] for multi-label classification.

References

1. Barnes, T.: Novel derivation and application of skill matrices: The Q-matrix method. Handbook on educational data mining pp. 159–172 (2010)
2. Chiu, C.Y.: Statistical refinement of the Q-matrix in cognitive diagnosis. Applied Psychological Measurement 37(8), 598–618 (2013)
3. Chiu, C.Y., Douglas, J.: A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. Journal of Classification 30(2), 225–250 (2013)
4. Desmarais, M., Beheshti, B., Xu, P.: The refinement of a Q-matrix: Assessing methods to validate tasks to skills mapping. In: Educational Data Mining 2014 (2014)
5. Desmarais, M.C.: Mapping question items to skills with non-negative matrix factorization. ACM SIGKDD Explorations Newsletter 13(2), 30–36 (2012)
6. Desmarais, M.C., Naceur, R.: A matrix factorization method for mapping items to skills and for enhancing expert-based q-matrices. In: Artificial Intelligence in Education. pp. 441–450. Springer (2013)
7. Desmarais, M.C., Xu, P., Beheshti, B.: Combining techniques to refine item to skills Q-matrices with a partition tree. In: Educational Data Mining 2015 (2015)
8. Grigorios Tsoumakas, Ioannis Katakis, I.V.: Data Mining and Knowledge Discovery Handbook. Springer (2010)
9. Henson, R.A., Templin, J.L., Willse, J.T.: Defining a family of cognitive diagnosis models using log-linear models with latent variables. Psychometrika 74(2), 191–210 (2009)
10. Nižnan, J., Pelánek, R., Řihák, J.: Mapping problems to skills combining expert opinion and student data. In: Mathematical and Engineering Methods in Computer Science, pp. 113–124. Springer (2014)
11. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Machine learning 85(3), 333–359 (2011)
12. Robitzsch, A., Kiefer, T., George, A.C., Uenlue, A.: CDM: Cognitive Diagnosis Modeling (2015), http://CRAN.R-project.org/package=CDM, r package version 4.5-0
13. Tatsuoka, K.K.: Rule space: An approach for dealing with misconceptions based on item response theory. Journal of educational measurement 20(4), 345–354 (1983)
14. de la Torre, J.: An empirically based method of Q-matrix validation for the dina model: Development and applications. Journal of educational measurement 45(4), 343–362 (2008)
15. de la Torre, J.: Dina model and parameter estimation: A didactic. Journal of Educational and Behavioral Statistics 34(1), 115–130 (2009)
16. Tsoumakas, G., Katakis, I., Vlahavas, I.: Random k-labelsets for multi-label classification. IEEE Transactions on Knowledge and Data Engineering 23(7), 1079–1089 (2011)

17. Wang, S., Douglas, J.: Consistency of nonparametric classification in cognitive diagnosis. Psychometrika 80(1), 85–100 (2015)

All links were last followed on