Tracking Knowledge Proficiency of Students with Educational Priors

Yuying Chen¹, Qi Liu¹, Zhenya Huang¹, Le Wu², Enhong Chen¹*, Runze Wu¹, Yu Su³, Guoping Hu⁴

¹ University of Science and Technology of China  
² Hefei University of Technology  
³ Anhui University  
⁴ iFLYTEK Research
Outline

- Background and Related Work
- Problem Statement
- Methodology
- Experiments
- Conclusion
Background

- **Traditional teaching method**
  - **Classroom Teaching**
    - The teacher’s energy is limited.
    - The same learning strategy, same exercises, impersonality.
  - **Extracurricular Tutorials**
    - Teaching quality is difficult to guarantee
    - A higher cost
Background

- E-Learning (Online learning)
  - Knewton
  - Cognitive Tutor
  - etc
Education Service Systems

Various online tutoring systems allow students to learn and do exercises individually.
Related work-static

- **IRT**
  \[
  P(X_{ij} = 1|\theta_j) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta_j - b_i)]}
  \]

- **DINA**
  \[
  P_j(\alpha_i) = P(X_{ij} = 1|\alpha_i) = g_j^{1 - \eta_{ij}}(1 - s_j)^{\eta_{ij}}.
  \]

- **PMF**
  \[
  p(R|U, V, \sigma^2) = \prod_{i=1}^{N} \prod_{j=1}^{M} \mathcal{N}(R_{ij}|U_i^T V_j, \sigma^2)^{I_{ij}}
  \]

they are only good at predicting student’s proficiency from a **static perspective.**
Related work - dynamic

- **LFA - one-dimensional**
  \[
  m(i, j \in KCs, n) = \alpha_i + \sum_{j \in KCs} (\beta_j + \gamma_j n_{i,j})
  \]
  \[
  p(m) = \frac{1}{1 + e^{-m}}
  \]

- **BKT - binary entities**
Outline

- Background and Related Work
- Problem Statement
- Methodology
- Experiments
- Conclusion
Motivation

- **Problem:** How to track students’ knowledge proficiency over time. (TKP task)?
- **Opportunity**
  - Widely use of Intelligent tutoring system
  - Record exercises logs and Q-matrix
  - Educational Priors
- **Focus on Math problem**

\[ \text{AB} = 3 \text{cm} \quad \text{AC} = 4 \text{cm} \]

\[ S_{\text{shadow}} = ? \text{ cm}^2 \]
Problem Statement

- Given the students’ response tensor $R$ and $Q$-matrix labelled by educational experts
- Our goal is two-fold:
  - Modeling the change of students’ knowledge proficiency from time 1 to $T$.
  - Predicting students’ knowledge proficiency and responses in time $T + 1$.
- Challenge:
  1. How to get a student’s knowledge proficiency?
  2. How to explain the change of knowledge proficiency over time?
A toy example

- A showcase of KPD task on mathematical exercises related to the knowledge points of Function and Inequality
Outline

- Background and Related Work
- Problem Statement
- **Methodology**
  - Probabilistic Modeling with Priors
  - Model Learning and Prediction
- Experiments
- Conclusion
KPT: a supervised way

Modeling Stage
- Partial order
- Forget
- Learn

Predicting Stage
- Input: $U^1, V$
- Output: $U^{T+1}, R^{T+1}$
KPT model

- **Probabilistic Modeling with Priors**
  - for each student and each exercise, we model the response tensor $R$ as:

  $$p(R|U, V, b) = \prod_{t=1}^{T} \prod_{i=1}^{N} \prod_{j=1}^{M} [\mathcal{N}(R_{ij}^t | \langle U_i^t, V_j \rangle - b_j, \sigma_R^2)]^{I_{ij}}_{t} ,$$

  where
  - $U_i^t \in \mathbb{R}^{K \times 1}$ is the knowledge proficiency of student $i$
  - $V \in \mathbb{R}^{M \times K}$ denotes the relationship between exercises and knowledge points

- How to establish the corresponding relationship between students, exercises and knowledge points?
Modeling V with the Q-matrix prior

- **Q-matrix**
  - depicts the knowledge points of the exercises
  - each row denotes an exercise
  - each column stands for a knowledge point.

<table>
<thead>
<tr>
<th></th>
<th>Function</th>
<th>Solid Geometry</th>
<th>Arithmetic Progression</th>
<th>Inequation</th>
</tr>
</thead>
<tbody>
<tr>
<td>exercise1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>exercise2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>exercise3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>exercise4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- The sparsity with the binary entities does not fit probabilistic modeling well.
Modeling V with the Q-matrix prior

- for exercise j, if a knowledge point q is marked as 1, then we assume that q is more relevant to exercise j than p with mark 0.

\[ q >_j^+ p, \text{ if } Q_{jq} = 1 \text{ and } Q_{jp} = 0. \]

- After that, we can transform the original Q-matrix into a set of comparability by:

\[ D_T \in \mathbb{R}^{M \times K \times K} \text{ by: } D_T = \{(j, q, p) | q >_j^+ p\}. \]
Modeling $V$ with the Q-matrix prior

- we define the probability that exercise $j$ is more relevant to knowledge point $q$ than knowledge point $p$ as:
  \[
p(q >_j^+ p|V_j) = \frac{1}{1 + e^{-(V_{jq} - V_{jp})}}.
\]  

- the log of the posterior distribution
  \[
  \ln p(V|D_T) = \ln \prod_{(j,q,p) \in D_T} p(>^+_j | V)p(V)
  = \sum_{j=1}^{M} \sum_{q=1}^{K} \sum_{p=1}^{K} I(q >^+_j p) \ln \frac{1}{1 + e^{-(V_{jq} - V_{jp})}} - \frac{1}{2\sigma_V^2} \|V\|_F^2.
  \]
we assume a student’s current knowledge proficiency is mainly influenced by two underlying reasons:

- She forgets her previous knowledge proficiency over time.
- The more exercises she does, the higher level of related knowledge proficiency she will get.

We model the two effects of each student’s knowledge proficiency in time window $t = 2; 3; \ldots; T$ as:

$$p(U_i^t) = \mathcal{N}(U_i^t | \tilde{U}_i^t, \sigma_U^2 I), \quad \text{where} \quad \tilde{U}_i^t = \{\tilde{U}_{i1}^t, \tilde{U}_{i2}^t \ldots \tilde{U}_{iK}^t\}$$

$$\tilde{U}_{ik}^t = (1 - \alpha_i) f^t(*) + \alpha_i l^t(*), \quad \text{s.t.} \quad 0 \leq \alpha_i \leq 1,$$  

(forgetting) (learning)
Modeling U with learning theories.

- $f^t(*)$ depicts the decline of knowledge over time:
  $$f^t(*) = U_{ik}^{t-1}e^{-\frac{\Delta t}{S}}$$
  - $\Delta t$ is the time interval
  - $S$ denotes the strength of memory.

- $l^t(*)$ captures the growth of knowledge with exercises:
  $$l^t(*) = U_{ik}^{t-1} \frac{Df_k^t}{f_k^t + r}$$
  - $f_k^t$ denotes the frequency of knowledge $k$
  - $r$ and $D$ control the magnitude and multiplier of growth respectively.
Model Learning and Prediction

- graphical representation of the proposed latent model
our goal is to learn the parameters $\Phi = [U, V, \alpha, b]$

- Particularly, the posterior distribution over $\Phi$ is:
  
  \[ p(U, V, \alpha, b|R, D_T) \propto p(R|U, V, \alpha, b) \times p(U|\sigma^2_U, \sigma^2_{U_1}) \times p(V|D_T). \]

- Maximizing the log posterior of the above equation is equivalent to minimize the following objective:

\[
\min_{\Phi} E(\Phi) = \frac{1}{2} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{M} I_{ij}^t \left[ \hat{R}_{ij}^t - R_{ij}^t \right]^2 \\
- \lambda_P \sum_{j=1}^{M} \sum_{q=1}^{K} \sum_{p=1}^{K} I(q > \hat{p}_j) \ln \frac{1}{1 + e^{-(V_{jq} - V_{jp})}} \\
+ \lambda_U \sum_{t=2}^{T} \sum_{i=1}^{N} \|U_i^t - U_i^t\|_F^2 \\
+ \frac{\lambda_{U1}}{2} \sum_{i=1}^{N} \|U_i^1\|_F^2 , \quad (13)
\]
With students’ knowledge proficiency $U_1^1, U_2^2, ..., U_T^T$ and related parameters, students’ responses and knowledge proficiency in the next time can be calculated as:

$$U_i^{(T+1)} = \{U_i^{(T+1)}_1, U_i^{(T+1)}_2, ..., U_i^{(T+1)}_K\},$$

$$U_{ik}^{(T+1)} \approx (1 - \alpha_i)U_{ik}^T e^{-\frac{\Delta(T)}{S}} + \alpha_i U_{ik}^T \frac{M f_k^{T+1}}{f_k^{T+1} + r},$$

$$\hat{R}_{ij}^{(T+1)} \approx \langle U_i^{(T+1)}, V_j \rangle - b_j.$$
Outline

- Background and Related Work
- Problem Statement
- Methodology
- Experiments
- Conclusion
Experiments

- **Dataset**
  - *Two private datasets* which are collected from daily exercise records of high school students
  - *ASSIST* is a public dataset *Assistments*\(^1\) 2009-2010 “Non-skill builder”

<table>
<thead>
<tr>
<th>Table 4: The statistics of the three datasets.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dataset</strong></td>
</tr>
<tr>
<td>Training scores logs</td>
</tr>
<tr>
<td>Testing scores logs</td>
</tr>
<tr>
<td>Students</td>
</tr>
<tr>
<td>Exercises</td>
</tr>
<tr>
<td>Time windows</td>
</tr>
<tr>
<td>Knowledge points</td>
</tr>
<tr>
<td>Average knowledge points of each exercise</td>
</tr>
</tbody>
</table>
Evaluations

- Two evaluations:
  - evaluate on Students’ Responses Prediction.
    - proved the **rationality of three priors for** prediction accuracy
  - evaluate on Knowledge Proficiency Diagnosis.
    - proved that the **effectiveness** of associating each exercise and student with a **knowledge vector** in the same knowledge space.
Evaluations on Students’ Responses Prediction.

- **Evaluation Metrics**
  - For Scores prediction task performance
    - RMSE, MAE
  - baselines:
    - IRT
    - DINA
    - PMF
    - LFA
    - BKT
    - QMIRT (MIRT+partial order)
    - QPMF (PMF+Partial order)

KPT performs best on all three datasets.
Evaluations on Knowledge Proficiency Diagnosis

- For Knowledge Proficiency Diagnosis
  - DOA-of each specific knowledge point $k$
    \[
    DOA(k) = \sum_{j=1}^{M} \sum_{u_1=1}^{N} \sum_{u_2=1}^{N} \frac{\delta (U_{u_1k}^T - U_{u_2k}^T) \cap \delta (R_{u_2j}^T - R_{u_1j}^T)}{\delta (U_{u_1k}^T - U_{u_1k}^T)},
    \]
  - DOA-average of all knowledge points
    \[
    DOA - \text{Avg} = \frac{1}{K} \sum_{k=1}^{K} DOA(k)
    \]

- baselines:
  - DINA
  - BKT
  - QMIRT (MIRT+partial order)
  - QPMF (PMF+Partial order)
KPT performs best on KPD task for all knowledge points, followed by QPMF and QIRT, which indicates that the educational prior of Q-matrix does effectively.
The diagnosis results of a student on six knowledge points at three particular time in Math2.

It clearly demonstrated the explanatory power of our proposed KPT model.
Outline

- Background and Related Work
- Problem Statement
- Methodology
- Experiments
- Conclusion
Conclusion

- **Problem**: track students’ knowledge proficiency mastery over time
- **Method**: probabilistic model with three educational priors

**Contributions:**
- We designed an *explanatory* probabilistic KPT model for solving the TKP task
- We associated each exercise with a knowledge vector with the $Q$-matrix prior.
- We embedded the *Learning curve* and *Forgetting curve* as priors to capture the change of each student’s proficiency over time.
Future Work

- First, we will consider to combine more kinds’ of users’ behaviors (e.g., reading records) for the TKP task.

- Second, as students may learn difficult knowledge points (e.g., Function) after some basic ones (e.g., Set), it is interesting to take this kind of knowledge relationship into account for TKP.
Acknowledgements

- We thank for:
  - the SIGIR Travel Award
    - url: http://sigir.org/general-information/travel-grants/
  - the SIGWEB and US NSF Travel Award
    - url: https://cmt3.research.microsoft.com/CIKMTA2017
Thanks!

Reporter: Yuying Chen
cyy33222@mail.ustc.edu.cn