Learning from Learning Curves: Item Response Theory & Learning Factors Analysis

Ken Koedinger
Human-Computer Interaction Institute
Carnegie Mellon University

Cognitive Tutor Technology
Use cognitive model to individualize instruction

- Cognitive Model: A system that can solve problems in the various ways students can

  \[
  3(2x - 5) = 9 \\
  \text{If goal is solve } a(bx+c) = d \\
  \text{Then rewrite as } abx + ac = d
  \]

  \[
  \text{Hint message: "Distribute } a \text{ across the parentheses."}
  \]

  \[
  \text{Known? = 85% chance}
  \]

  \[
  6x - 15 = 9 \\
  2x - 5 = 3 \\
  6x - 5 = 9
  \]

- Model Tracing: Follows student through their individual approach to a problem -> context-sensitive instruction

- Knowledge Tracing: Assesses student's knowledge growth -> individualized activity selection and pacing

Cognitive Model Discovery

- Traditional Cognitive Task Analysis
  - Interview experts, think alouds, DFA

- Result: cognitive model of student knowledge
  - Cognitive model drives ITS behaviors & instructional design decisions

- Key goal for Educational Data Mining
  - Improve Cognitive Task Analysis
  - Use student data from initial tutor
  - Employ machine learning & statistics to discover better cognitive models
Overview

- Using learning curves to evaluate cognitive models
- Statistical models of student performance & learning
  - Example of improving tutor
  - Comparison to other Psychometric models
- Using Learning Factors Analysis to discover better cognitive models
- Educational Data Mining research challenges

Production Rule Analysis

Evidence for Production Rule as an appropriate unit of knowledge acquisition

Using learning curves to evaluate a cognitive model

- Lisp Tutor Model
  - Learning curves used to validate cognitive model
  - Fit better when organized by knowledge components (productions) rather than surface forms (programming language terms)
- But, curves not smooth for some production rules
  - “Blips” in learning curves indicate the knowledge representation may not be right
  - Let me illustrate …
Curve for “Declare Parameter” production rule

- How are steps with blips different from others?
- What’s the unique feature or factor explaining these blips?

Can modify cognitive model using unique factor present at “blips”

- Blips occur when to-be-written program has 2 parameters
- Split Declare-Parameter by parameter-number factor:
  - Declare-first-parameter
  - Declare-second-parameter

Can learning curve analysis be automated?

- Manual learning curve analysis
  - Identify “blips” in learning curve visualization
  - Manually create a new model
  - Qualitative judgment of fit

- Toward automatic learning curve analysis
  - Blips as deviations from statistical model
  - Propose alternative cognitive models
  - Evaluate cognitive model using prediction accuracy statistics

Overview

- Using learning curves to evaluate cognitive models
- Statistical models of student performance & learning
  - Example of improving tutor
  - Comparison to other Psychometric models
- Using Learning Factors Analysis to discover better cognitive models
- Educational Data Mining research challenges
Representing Knowledge Components as factors of items

- Problem: How to represent KC model?
- Solution: Q-Matrix (Tatsuoka, 1983)

<table>
<thead>
<tr>
<th>Item</th>
<th>KCs: Add</th>
<th>Sub</th>
<th>Mul</th>
<th>Div</th>
</tr>
</thead>
<tbody>
<tr>
<td>2*8</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2*8 - 3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- Single KC item = when a row has one 1
- Multi-KC item = when a row has many 1’s

Q matrix is a bridge between a symbolic cognitive model & a statistical model

Additive Factors Model Assumptions

- Logistic regression to fit learning curves (Draney, Wilson, Pirolli, 1995)
- Assumptions about knowledge components (KCs) & students
  - Different students may initially know more or less
  - Students generally learn at the same rate
  - Some KCs are initially easier than others
  - Some KCs are easier to learn than others

- These assumptions are reflected in a statistical model
  - Intercept parameters for each student
  - Intercept & slope parameters for each KC
  - Slope = for every practice opportunity there is an increase in predicted performance

Simple Statistical Model of Performance & Learning

- Problem: How to predict student responses from model?
- Solution: Additive Factor Model
  - $i$ students, $j$ problems/items, $k$ knowledge components (KCs)

$$\log{\frac{p_{ij}}{1-p_{ij}}} = \theta_i + \sum_{k} \beta_{k}Q_{kj} + \sum_{k} Q_{kj}(\gamma_{k}T_{ik})$$

Model parameters:

- Student intercept
- KC intercept
- KC slope

Area Unit of Geometry Cognitive Tutor

- Original cognitive model in tutor:
  - 15 skills:
    - Circle-area
    - Circle-circumference
    - Circle-diameter
    - Circle-radius
    - Compose-by-addition
    - Compose-by-multiplication
  - Parallelogram-area
  - Parallelogram-side
  - Pentagon-area
  - Pentagon-side
  - Trapezoid-area
  - Trapezoid-base
  - Trapezoid-height
  - Triangle-area
  - Triangle-side
Log Data Input to AFM

- **Items = steps in tutors with step-based feedback**
- **Q-matrix in single column: works for single KC items**
- **Opportunities:** Student has had to learn KC

<table>
<thead>
<tr>
<th>Student</th>
<th>Step (Item)</th>
<th>KC</th>
<th>Opportunity</th>
<th>Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>p1s1</td>
<td>Circle-area</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>p2s1</td>
<td>Circle-area</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>p2s2</td>
<td>Rectangle-area</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>p2s3</td>
<td>Compose-by-addition</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>p3s1</td>
<td>Circle-area</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

AFM Results for original KC model

- Higher intercept of skill -> easier skill
- Higher slope of skill -> faster students learn it

<table>
<thead>
<tr>
<th>Skill</th>
<th>Intercept</th>
<th>Slope</th>
<th>Avg Opportunities</th>
<th>Initial Probability</th>
<th>Avg Probability</th>
<th>Final Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallelogram-area</td>
<td>2.14</td>
<td>-0.01</td>
<td>14.9</td>
<td>0.95</td>
<td>0.94</td>
<td>0.93</td>
</tr>
<tr>
<td>Pentagon-area</td>
<td>-2.16</td>
<td>0.45</td>
<td>4.3</td>
<td>0.2</td>
<td>0.63</td>
<td>0.64</td>
</tr>
</tbody>
</table>

- **Student Intercept:**
  - Student0: 1.38
  - Student1: 0.82
  - Student2: 0.21

- **Model Statistics:**
  - AIC: 3,950
  - BIC: 4,285
  - MAD: 0.083

**Overview**

- Using *learning curves* to evaluate cognitive models
- Statistical models of student performance & learning
  - Example of improving tutor
  - Comparison to other Psychometric models
- Using *Learning Factors Analysis* to discover better cognitive models
- Educational Data Mining research challenges

**Application: Use Statistical Model to improve tutor**

- Some KCs over-practiced, others under (Cen, Koedinger, Junker, 2007)

**AFM Results**

- **Initial error rate:**
  - 76% reduced to 40% after 6 times of practice
  - 12% reduced to 8% after 18 times of practice
“Close the loop” experiment

- **In vivo** experiment: New version of tutor with updated knowledge tracing parameters vs. prior version
- Reduced learning time by 20%, same robust learning gains
- Knowledge transfer: Carnegie Learning using approach for other tutor units

### Comparing to other psychometric models

- AFM adds a growth component to “LLTM” (Wilson & De Boeck)
  - LLTM is an “item explanatory” generalization of IRT or “Rasch”
  - “Person explanatory” models are related to factor analysis and

<table>
<thead>
<tr>
<th>Model</th>
<th>Person part</th>
<th>Item part</th>
<th>Random effect</th>
<th>Model type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rasch model</td>
<td>$\theta_p$</td>
<td>$-\beta_i$</td>
<td>$\epsilon_p \sim N(0, \sigma^2_p)$</td>
<td>Doubly descriptive</td>
</tr>
<tr>
<td>Latent reg</td>
<td>$\sum_{j=1}^J \theta_j Z_{pj} + \epsilon_p$</td>
<td>$-\beta_i$</td>
<td>$\epsilon_p \sim N(0, \sigma^2_p)$</td>
<td>Person explanatory</td>
</tr>
<tr>
<td>Rasch model</td>
<td>$\theta_p$</td>
<td>$-\sum_{k=1}^K \beta_k X_{ik}$</td>
<td>$\theta_p \sim N(0, \sigma^2_p)$</td>
<td>Item explanatory</td>
</tr>
<tr>
<td>LLTM</td>
<td>$\sum_{j=1}^J \theta_j Z_{pj} + \epsilon_p$</td>
<td>$-\sum_{k=1}^K \beta_k X_{ik}$</td>
<td>$\epsilon_p \sim N(0, \sigma^2_p)$</td>
<td>Doubly explanatory</td>
</tr>
</tbody>
</table>

Additive Factor Model (AFM) generalizes Item Response Theory (IRT)

- Instance of logistic regression
  - Example: In R use generalized linear regression with family=binomial
    - \( \text{glm}(\text{prob-correct} ~ \text{student} + \text{KC} + \text{KC:opportunity}, \text{family}=\text{binomial},...) \)
- Generalization of item response theory (IRT)
  - IRT simply has i student & j item parameters
    - \( \text{glm}(\text{prob-correct} ~ \text{student} + \text{item}, \text{family}=\text{binomial},...) \)
- AFM is different from IRT because:
  - It clusters items by knowledge components
  - It has an opportunity slope for each KC

Model Evaluation

- How to compare cognitive models?
  - A good model minimizes prediction risk by balancing fit with data & complexity (Wasserman 2005)
- Model-data fit metrics
  - Log likelihood, root mean squared error (RMSE), mean average deviation (MAD), area under curve (AUC), …
- Prediction metrics
  - BIC, AIC: Faster metrics add a penalty for # parameters
    - \( \text{BIC} = -2\log-likelihood + \text{numPar} \times \log(\text{numOb}) \)
  - Cross validation: Slower but better
    - Split data in training & test sets, optimize parameters with training set, apply fit metrics on test set
A good cognitive model produces a learning curve. Recall LISP tutor example above.

Without decomposition, using just a single “Geometry” skill, there is no smooth learning curve.

But with decomposition, 12 skills for area, there is a smooth learning curve.

Is this the correct or “best” cognitive model?

DataShop visualizations to aid “blip” detection:

- Many curves show a reasonable decline.
- Some do not → Opportunity to improve model!

Overview

- Using learning curves to evaluate cognitive models
- Statistical models of student performance & learning
  - Example of improving tutor
  - Comparison to other Psychometric models
- Using Learning Factors Analysis to discover better cognitive models
- Educational Data Mining research challenges
Learning Factors Analysis (LFA): A Tool for Cognitive Model Discovery

- LFA is a method for discovering & evaluating alternative cognitive models
  - Finds knowledge components that best predict student performance & learning transfer
- Inputs
  - Data: Student success on tasks in domain over time
  - Codes: Factors hypothesized to drive task difficulty & transfer
- Outputs
  - A rank ordering of most predictive cognitive models
  - Parameter estimates for each model

Learning Factors Analysis (LFA) draws from multiple disciplines

- Cognitive Psychology
  - Learning curve analysis (Corbett, et al 1995)
- Psychometrics & Statistics
  - Q Matrix & Rule Space (Tatsuoka 1983, Barnes 2005)
  - Item response learning model (Draney, et al., 1995)
  - Item response assessment models (DiBello, et al., 1995; Embretson, 1997; von Davier, 2005)
- Machine Learning & AI
  - Combinatorial search (Russell & Norvig, 2003)

Item Labeling & the “P Matrix”: Adding Alternative Factors

- How to improve existing cognitive model?
- Have experts look for difficulty factors that are candidates for new KCs. Put these in "P matrix"

Using P matrix to update Q matrix

- Create a new Q’ by using elements of P as arguments to operators
  - Add operator: Q’ = Q + P[,1]
  - Split operator: Q’ = Q[, 2] * P[,1]
LFA: KC Model Search

- How to find best model given Q and P matrices?
- Use best-first search algorithm (Russell & Norvig 2002)
  - Guided by a heuristic, such as BIC or AIC
- Do model selection within space of Q matrices

Steps:
1. Start from an initial “node” in search graph using given Q
2. Iteratively create new child nodes (Q’) by applying operators with arguments from P matrix
3. Employ heuristic (BIC of Q’) to rank each node
4. Select best node not yet expanded & go back to step 2

Example in Geometry of split based on factor in P matrix

<table>
<thead>
<tr>
<th>Original Q matrix</th>
<th>Factor in P matrix</th>
<th>After Splitting Circle-area by Embed</th>
<th>New Q matrix</th>
<th>Revised Opportunity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>Step</td>
<td>Skill</td>
<td>Opportunity</td>
<td>Embed</td>
</tr>
<tr>
<td>A</td>
<td>p1x1</td>
<td>Circle-area</td>
<td>0</td>
<td>alone</td>
</tr>
<tr>
<td>A</td>
<td>p2x1</td>
<td>Circle-area</td>
<td>1</td>
<td>embed</td>
</tr>
<tr>
<td>A</td>
<td>p2x2</td>
<td>Rectangle-area</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>p2x3</td>
<td>Compose-by-add</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>p3x1</td>
<td>Circle-area</td>
<td>2</td>
<td>alone</td>
</tr>
</tbody>
</table>

Example LFA Results: Applying splits to original model

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Splits: 3</td>
<td>Number of Splits: 3</td>
<td>Number of Splits: 2</td>
</tr>
<tr>
<td>2. Binary split circle-radius by repeat repeat</td>
<td>2. Binary split circle-radius by repeat repeat</td>
<td>2. Binary split circle-radius by repeat repeat</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Skills: 18</th>
<th>Number of Skills: 18</th>
<th>Number of Skills: 17</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIC: 4,248.86</td>
<td>BIC: 4,248.86</td>
<td>BIC: 4,251.07</td>
</tr>
</tbody>
</table>

- Common results:
  - Compose-by-multiplication split based on whether it was an area or a segment being multiplied
  - Circle-radius is split based on whether it is being done for the first time in a problem or is being repeated

Compose-by-multiplication KC examples

Composing Areas

Composing Segments

Tutor Design Implications 1

- LFA search suggests distinctions to address in instruction & assessment
  - With these new distinctions, tutor can
    - Generate hints better directed to specific student difficulties
    - Improve knowledge tracing & problem selection for better cognitive mastery
- Example: Consider Compose-by-multiplication before LFA

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>slope</th>
<th>Avg Practice Opportunities</th>
<th>Initial Probability</th>
<th>Avg Probability</th>
<th>Final Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>CM</td>
<td>-.15</td>
<td>.1</td>
<td>10.2</td>
<td>.65</td>
<td>.84</td>
<td>.92</td>
</tr>
<tr>
<td>CMarea</td>
<td>-.009</td>
<td>.17</td>
<td>9</td>
<td>.64</td>
<td>.86</td>
<td>.96</td>
</tr>
<tr>
<td>CMseg</td>
<td>-1.42</td>
<td>.48</td>
<td>1.9</td>
<td>.32</td>
<td>.54</td>
<td>.60</td>
</tr>
</tbody>
</table>

With final probability .92, many students are short of .95 mastery threshold.

Tutor Design Implications 2

- However, after split:
  - CM-area and CM-segment look quite different
    - CM-area is now above .95 mastery threshold (at .96)
    - But CM-segment is only at .60
    - Original model penalizes students who have key idea about composite areas (CM-area) -- some students solve more problems than needed
- Instructional redesign implications:
  - Change skillometer so CM-area & CM-segment are separately addressed
    - Set parameters appropriately -- CM-segment with a lower initial known value
  - Add more problems to allow for mastery of CM-segment
  - Add new hints specific to the CM-segment situation

Summary of Learning Factors Analysis (LFA)

- LFA combines statistics, human expertise, & combinatorial search to discover cognitive models
- Evaluates a single model in seconds, searches 100s of models in hours
  - Model statistics are meaningful
  - Improved models suggest tutor improvements
- Can currently be applied, by request, to any dataset in DataShop with at least two KC models
Mixed initiative human-machine discovery

1. Human
   - Hypothesize possible “learning factors” and code steps

2. Machine
   - Search over factors, report best models discovered

3. Human
   - Inspect results
   - If needed, propose new factors. Go to 2.
   - If good, modify tutor and test.

Human-machine discovery of new cognitive models

- Better models discovered in Geometry, Statistics, English, Physics

<table>
<thead>
<tr>
<th>Dataset name</th>
<th>Domain</th>
<th>Existing best BIC</th>
<th># of KCs</th>
<th>LFA Discovered BIC</th>
<th># of KCs</th>
<th>Improved BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geometry96-97</td>
<td>Geometry Area</td>
<td>5066</td>
<td>12</td>
<td>5548</td>
<td>10</td>
<td>1%</td>
</tr>
<tr>
<td>Hampton0506</td>
<td>Geometry Area</td>
<td>15047</td>
<td>18</td>
<td>12476</td>
<td>15</td>
<td>17%</td>
</tr>
<tr>
<td>Cog discovery</td>
<td>Geometry Area</td>
<td>31183</td>
<td>49</td>
<td>31109</td>
<td>29</td>
<td>0.2%</td>
</tr>
<tr>
<td>Statistics - Fall 2009</td>
<td>Statistics</td>
<td>3611</td>
<td>14</td>
<td>3454</td>
<td>8</td>
<td>4%</td>
</tr>
<tr>
<td>IWT S-E Study 2</td>
<td>English articles</td>
<td>7162</td>
<td>19</td>
<td>7068</td>
<td>11</td>
<td>1%</td>
</tr>
<tr>
<td>Physics 2009 Spring</td>
<td>Physics</td>
<td>27051</td>
<td>239</td>
<td>24917</td>
<td>14</td>
<td>8%</td>
</tr>
</tbody>
</table>

Open Research Questions: Technical

- What factors to consider? P matrix is hard to create
  - Enhancing human role: Data visualization strategies
  - Other techniques: Matrix factorization, LiFT
  - Other data: Do clustering on problem text
- Interpreting LFA output can be difficult
  - How to make interpretation easier?

=> Researcher can’t just “go by the numbers”
   1) Understand the domain, the tasks
   2) Get close to the data
Model search using DataShop: Human & machine improvements

- DataShop datasets w/ improved KC models:
  - Geometry Area (1996-1997), Geometry Area Hampton 2005-2006 Unit 34, …
- New KCs (learning factors) found using DataShop visualization tools
  - Learning curve, point tool, performance profiler
  - Example of human “feature engineering”
- New KC models also discovered by LFA
  - Research goal: Iterate between LFA & visualization to find increasingly better KC models

Detecting planning skills: Scaffolded vs. unscaffolded problems

- Scaffolded
  - Prompts are given for subgoals
- Un scaffolded
  - Prompts are not given for subgoals (initially)

Most curves “curve”, but if flat, then KC may be bad

Discovering a new knowledge component

- Each KC should have:
  1. smooth learning curve
  2. statistical evidence of learning
  3. even error rates across tasks
- Create new KCs by finding a feature common to hard tasks but missing in easy ones

<table>
<thead>
<tr>
<th>KC Name</th>
<th>Intercept</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>circle-area</td>
<td>0.76</td>
<td>0.74</td>
</tr>
<tr>
<td>composite-by-addition</td>
<td>0.8</td>
<td>0.72</td>
</tr>
<tr>
<td>pentagon-area</td>
<td>0.47</td>
<td>0.18</td>
</tr>
<tr>
<td>trapezoid-area</td>
<td>0.59</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Easy tasks do not require subgoals, hard tasks do? Not smooth

No learning

Uneven error rate
New model discovery: Split “compose” into 3 skills

- Hidden planning knowledge:
  - If you need to find the area of an irregular shape, then try to find the areas of regular shapes that make it up
- Redesign instruction in tutor
  - Design tasks that isolate the hidden planning skill
  - Given square & circle area, find leftover

3-way split in new model (green) better fits variability in error rates than original (blue)

Before unpacking compose-by-addition

After -- unpacked into subtract, decompose, remaining compose-by-addition

Automate human-machine strategies for “blip” detection

- Research goal: Automate low slope, non-low intercept, & high residual detection
- Uses:
  - speed up LFA search
  - point human coders to bad KCs
    - cluster harder vs. easier tasks
Developing & evaluating different learning curve models

- Many papers in Educational Data Mining (EDM) conference
  - Also in Knowledge Discovery & Data mining (KDD)
- Papers comparing knowledge tracing, AFM, PFA, CPFA, IFA
  - See papers by Pavlik, Beck, Chi …

Open Research Questions: Psychology of Learning

- Change AFM model assumptions
  - Is student learning rate really constant?
  - Does a Student x Opportunity interaction term improve fit?
  - What instructional conditions or student factors change rate?
- Is knowledge space “uni-dimensional”?
  - Does a Student x KC interaction term improve fit?
- Need different KC models for different students/conditions?
- Is learning curve an exponential or power law?
  - Long-standing debate, which has focused on “reaction time” not on error rate!
  - Compare use of Opportunity vs.Log(Opportunity)
- Other outcome variables: reaction time, assistance score
- Other predictors: Opportunities => Time per instructional event;
  Kinds of opportunities: Successes, failures, hints, gamed steps, …

Open Research Questions: Instructional Improvement

- Do LFA results generalize across data sets?
  - Is AIC or BIC a good estimate for cross-validation results?
  - Does a model discovered with one year’s tutor data generalize to a next year?
  - Does model discovery work for ed games, other domains?
- Use learning curves to compare instructional conditions in experiments
- Need more “close the loop” experiments
  - EDM => better model => better tutor => better student learning

EDM => better model => better tutor => better student learning