Measuring Similarity of Educational Items Using Data on Learners' Performance

Jiří Řihák, Radek Pelánek

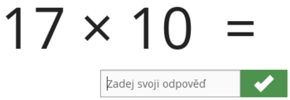
Masaryk University Brno

Adaptive
Learning
Research group, Masaryk university Brno

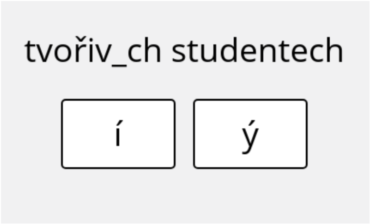
Adaptive learning

Adaptive practice systems

- **items** — simple questions
- practice — rapid sequence of items


17 × 10 =

Zadej svoji odpověď


tvořiv_ch studentech

í ý

Motivation

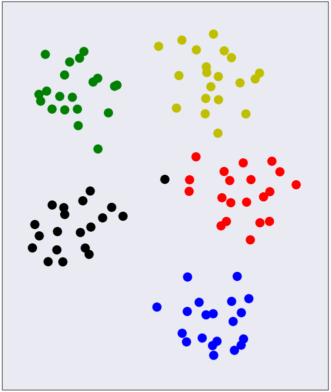
Large pool of items

- How to organize these items?
- What knowledge components should be used?
- Are there some anomalies?
- . . .

Motivation

Large pool of items

- **clustering**
- visualization
- outlier detection
- . . .

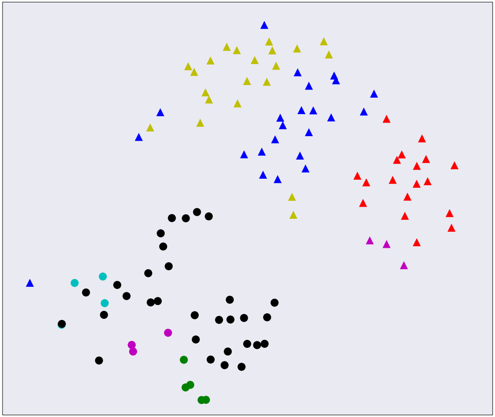


Adaptive Learning

Motivation

Large pool of items

- clustering
- **visualization**
- outlier detection
- . . .

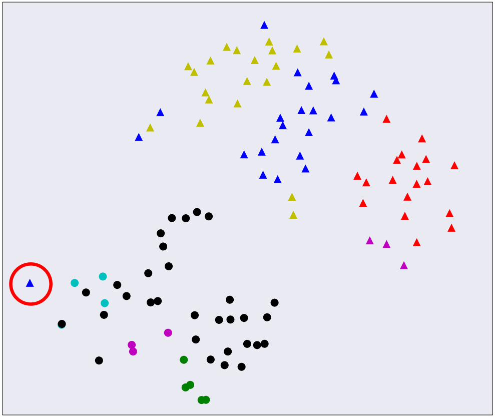


Adaptive Learning

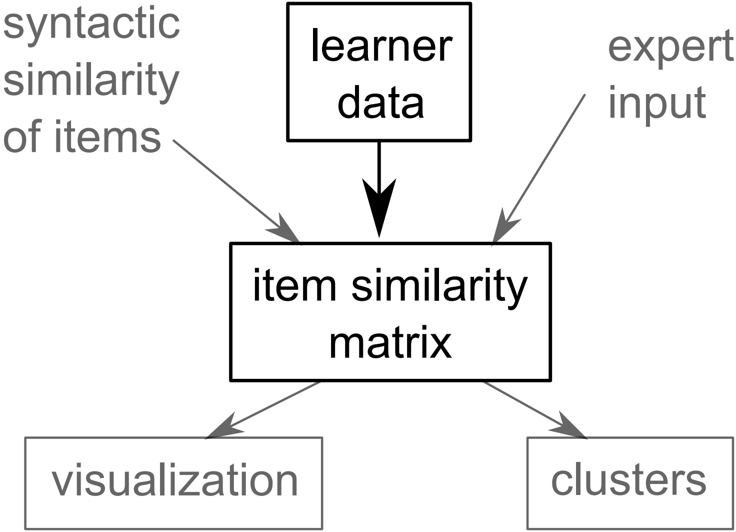
Motivation

Large pool of items

- clustering
- visualization
- **outlier detection**
- ...



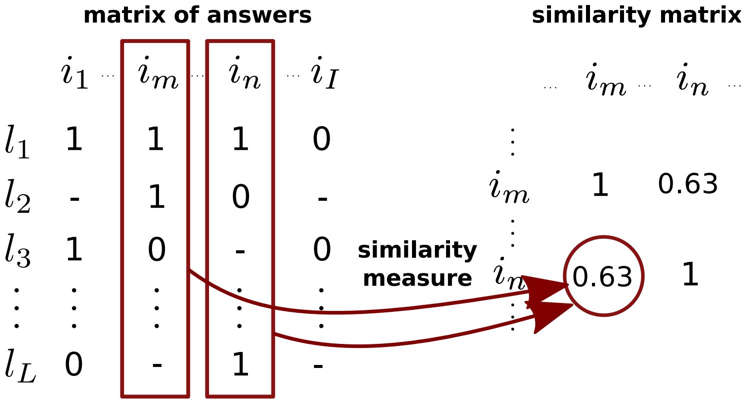
General approach



Research questions

- What similarity measures is suitable for EDM?
- How much data we need?
- How to combine more types of learner data?

Similarity measures



matrix of answers

$$\begin{array}{cccccccc}
 & i_1 & \cdots & i_m & \cdots & i_n & \cdots & i_I \\
l_1 & 1 & & 1 & & 1 & & 0 \\
l_2 & - & & 1 & & 0 & & - \\
l_3 & 1 & & 0 & & - & & 0 \\
\vdots & \vdots & & \vdots & & \vdots & & \vdots \\
l_L & 0 & & - & & 1 & & -
\end{array}$$

similarity measure

similarity matrix

$$\begin{array}{cccccc}
 & \cdots & i_m & \cdots & i_n & \cdots \\
\vdots & & & & & \\
i_m & & 1 & & 0.63 & \\
\vdots & & & & & \\
i_n & & \boxed{0.63} & & 1 & \\
\vdots & & & & &
\end{array}$$

Similarity measures

binary data

- 1 — correct
- 0 — incorrect
- input can be simplified:

		item i	
		incorrect	correct
item j	incorrect	a	b
	correct	c	d

Similarity measures

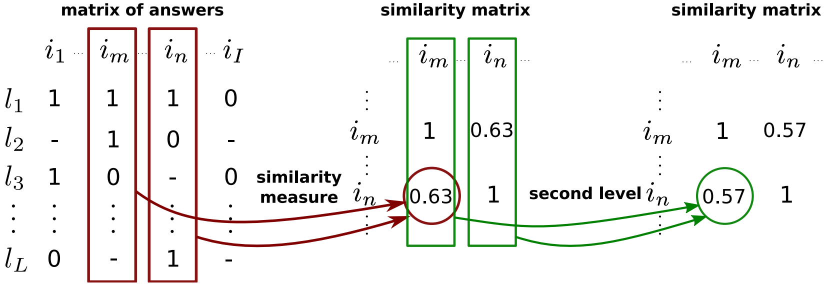
Yule	$S_y = (ad - bc)/(ad + bc)$
Pearson	$S_p = (ad - bc)/\sqrt{(a+b)(a+c)(b+d)(c+d)}$
Cohen	$S_c = (P_o - P_e)/(1 - P_e)$ $P_o = (a+d)/n$ $P_e = ((a+b)(a+c) + (b+d)(c+d))/n^2$
Sokal	$S_s = (a+d)/(a+b+c+d)$
Jaccard	$S_j = a/(a+b+c)$
Ochiai	$S_o = a/\sqrt{(a+b)(a+c)}$

Similarity measures

Yule	
Pearson	(Matthews) correlation coefficient, Phi coefficient
Cohen	Cohen's kappa

Sokal	Sokal-Michener, simple matching
Jaccard	
Ochiai	cosine similarity

Second level of similarity



matrix of answers

$$\begin{array}{c|ccccccc} & i_1 & \cdots & i_m & \cdots & i_n & \cdots & i_I \\ \hline l_1 & 1 & & 1 & & 1 & & 0 \\ l_2 & - & & 1 & & 0 & & - \\ l_3 & 1 & & 0 & & - & & 0 \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ l_L & 0 & & - & & 1 & & - \end{array}$$

similarity measure

similarity matrix

$$\begin{array}{c|ccccc} & \cdots & i_m & \cdots & i_n & \cdots \\ \hline \vdots & & & & & \\ i_m & & 1 & & 0.63 & \\ \vdots & & & & & \\ i_n & & 0.63 & & 1 & \\ \vdots & & & & & \end{array}$$

second level

similarity matrix

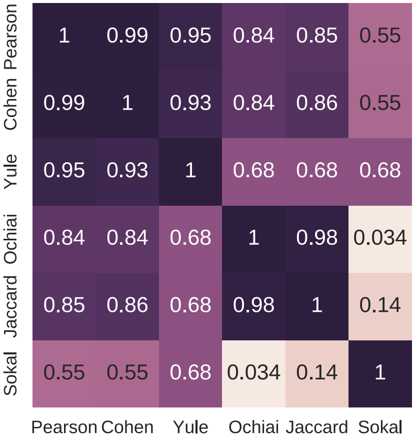
$$\begin{array}{c|ccccc} & \cdots & i_m & \cdots & i_n & \cdots \\ \hline \vdots & & & & & \\ i_m & & 1 & & 0.57 & \\ \vdots & & & & & \\ i_n & & 0.57 & & 1 & \\ \vdots & & & & & \end{array}$$

Second level of similarity

- 2 items are close if they are *similarly* close to other items
- more information used
- noise reduction
- necessary for some follow up algorithms

Evaluation - correlation of measures

- Cohen - **Pearson**
- Ochiai - **Jaccard**
- Yule
- Sokal - the most different



Czech 1 (adjectives)

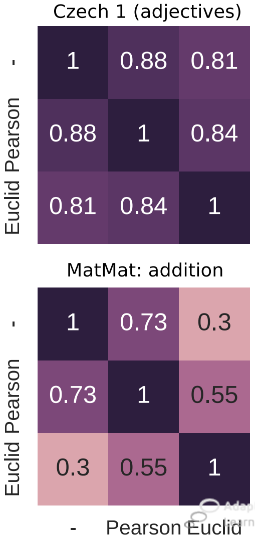
	Pearson	Cohen	Yule	Ochiai	Jaccard	Sokal
Pearson	1	0.99	0.95	0.84	0.85	0.55
Cohen	0.99	1	0.93	0.84	0.86	0.55
Yule	0.95	0.93	1	0.68	0.68	0.68
Ochiai	0.84	0.84	0.68	1	0.98	0.034
Jaccard	0.85	0.86	0.68	0.98	1	0.14
Sokal	0.55	0.55	0.68	0.034	0.14	1

Evaluation - correlation of measures

- Cohen - **Pearson**
- Ochiai - **Jaccard**
- Yule
- Sokal - the most different

Second level of similarity

- brings change
- larger for smaller datasets



Czech 1 (adjectives)

	-	Pearson	Euclid
-	1	0.88	0.81
Pearson	0.88	1	0.84
Euclid	0.81	0.84	1

MatMat: addition

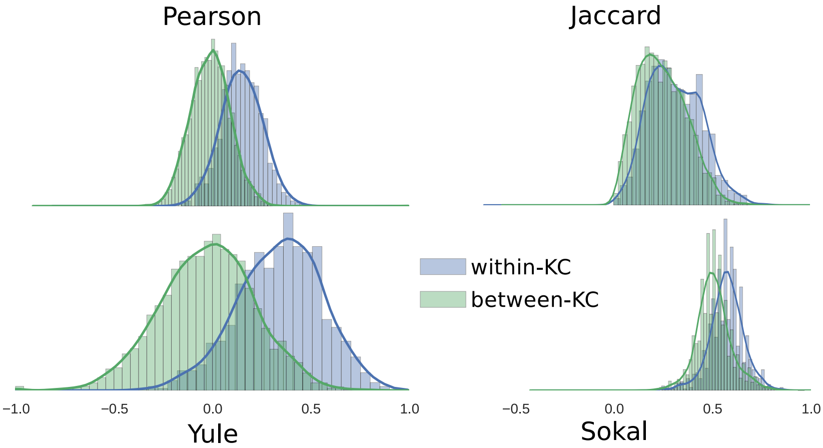
	-	Pearson	Euclid
-	1	0.73	0.3
Pearson	0.73	1	0.55
Euclid	0.3	0.55	1

Simulated data

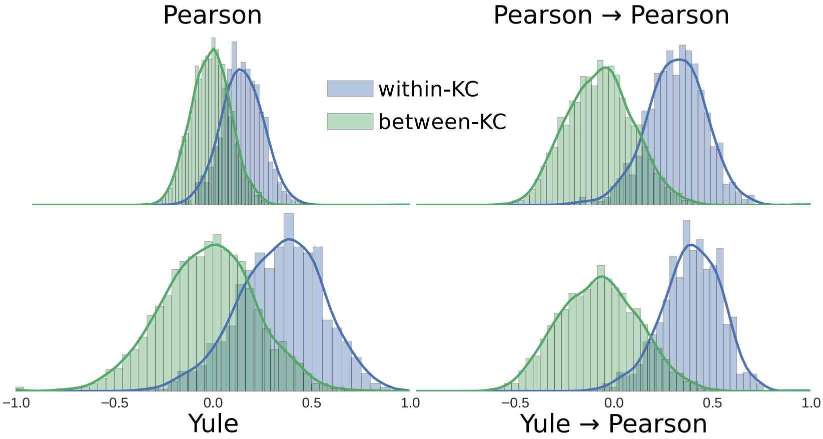
Simulated data

- we know *right answer*
- logistic model
  - learners have skills
  - items have difficulty
- typical setting
  - 100 learners
  - 5 knowledge components
  - 20 items per KC

Evaluation


Pearson
Jaccard
within-KC
between-KC
−1.0
−0.5
0.0
0.5
1.0
Yule
−0.5
0.0
0.5
1.0
Sokal

Evaluation


Pearson
Pearson → Pearson
within-KC
between-KC
−1.0
−0.5
0.0
0.5
1.0
−0.5
0.0
0.5
1.0
Yule
Yule → Pearson

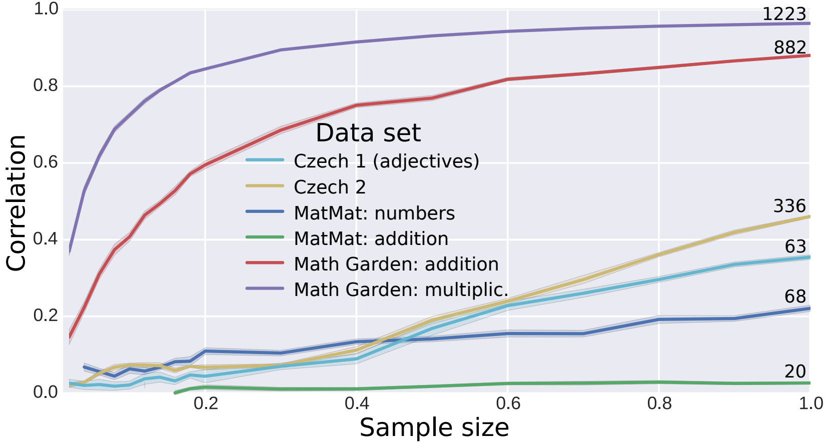
Evaluation - clustering

	Czech adjectives	100L 5KC	200L 5KC
Pearson	0.32 ± 0.02	0.48 ± 0.05	0.84 ± 0.05
Jaccard	0.31 ± 0.03	0.15 ± 0.04	0.29 ± 0.08
Yule	0.31 ± 0.03	0.43 ± 0.05	0.77 ± 0.07
Sokal	0.15 ± 0.06	0.18 ± 0.03	0.25 ± 0.05
Pearson → Euclid	**0.43** ± 0.01	**0.80** ± 0.06	**0.98** ± 0.01
Yule → Euclid	0.32 ± 0.02	0.65 ± 0.07	0.94 ± 0.04
Pearson → Pearson	0.41 ± 0.03	0.73 ± 0.06	0.96 ± 0.02
Yule → Pearson	0.32 ± 0.03	0.72 ± 0.06	0.97 ± 0.02

Do We Have Enough Data?

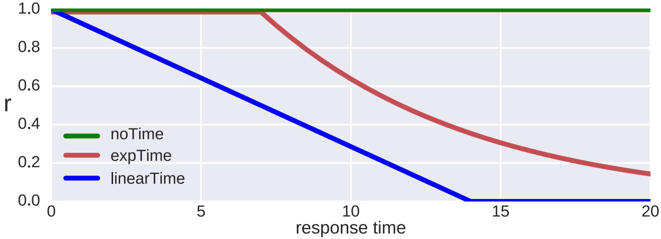
- stability of results
- split data to two halves
- how similarity measures correlate on these halves?

Do We Have Enough Data?



Response times

- additional information
- correctness and response time to one measure of success
- response: $0/1 \rightarrow [0, 1]$

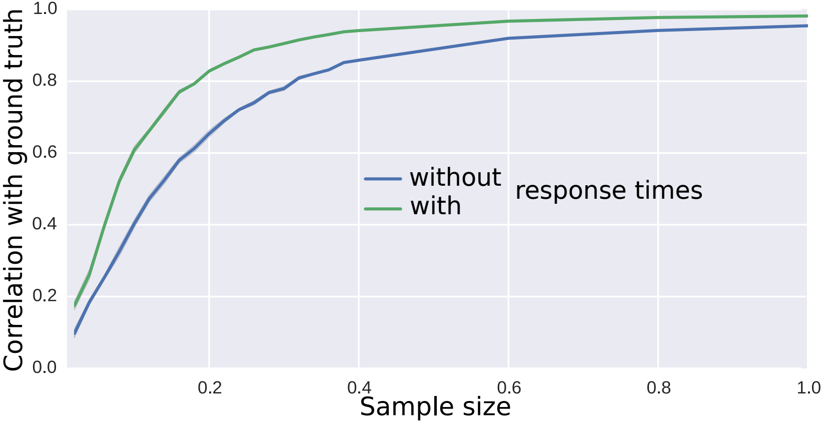


Adaptive Learning

Respose times - Math Garden

- Math Garden - large datasets: $\sim 1M$ of answer on 30 items
- small impact of time information - correlation > 0.9
- but what with smaller datasets?

Respose times - MathGarden



Conclusion

- **Pearson metric is a good default**
- Pearson, Cohen and Yule are good
- second level improve results
- we should check that we have sufficient data
- response time can help with small datasets