Exceptional Subitizing Patterns: Exploring Mathematical Abilities of Finnish Primary School Children with Piecewise Linear Regression

Seminário 3 – Aprendizado Descritivo

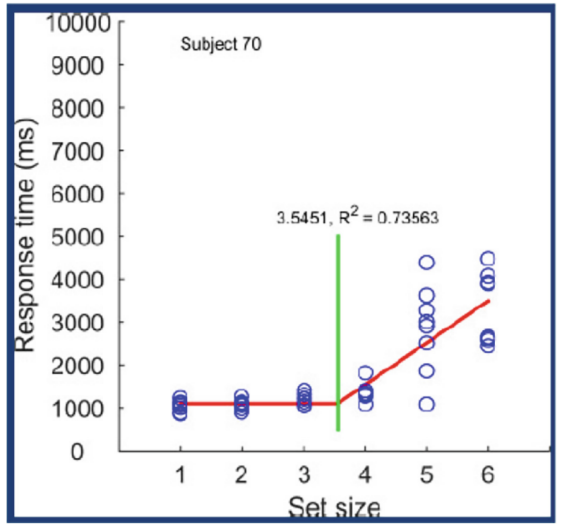
2025.1

Historiador

GABRIEL TONIONI DUARTE
ENILDA ALVES COELHO
FÁBIO CÉSAR MARRA FILHO
IASMIN CORREA ARAÚJO
JOSÉ VINÍCIUS DE LIMA MASSARICO
LARISSA DUARTE SANTANA
MARCELO LOMMEZ RODRIGUES DE JESUS

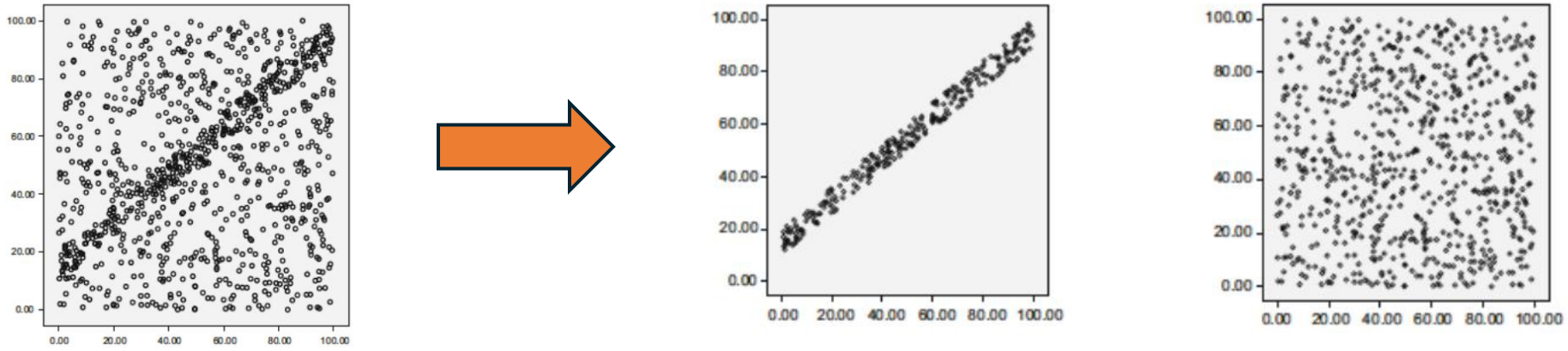
Objetivo e contexto

- Identificar **padrões atípicos de subitização** em crianças do ensino fundamental, fazendo a mineração de modelos excepcionais com regressão linear segmentada.
- Subitização: habilidade de identificar quantos elementos estão em um pequeno grupo sem a necessidade de contar.
  - A incapacidade de subitizar está associada à discalculia.
- Dados fornecidos pelo estudo FUNA



Técnicas relacionadas

- **Descoberta de subgrupos:** encontrar "subgrupos que sejam tão grandes quanto possível e apresentem as características estatísticas (de distribuição) mais incomuns em relação à propriedade de interesse" (variável alvo).
- **Mineração de modelos excepcionais (EMM):** ainda queremos descobrir subgrupos, mas agora olhando para a correlação de mais de uma variável alvo.



Trabalhos relacionados

- [19]: Uso de linear target models em EMM.
- [26] e [23]: QMs (métricas de qualidade): baseadas em log-verossimilhança.
- [23] e [3]: Atributos variando com o tempo, em contextos de flutuações de glicose e aplicações de financiamento.
- [27] e [11]: Abordagens: a primeira (RSD) é baseada em proposições, utilizando Prolog. Já a segunda é descrita como "abordagem simples para agrupamento de atributos".

Formalizador

DANIEL SCHLICKMANN BASTOS

GABRIEL CASTELO BRANCO ROCHA

GUILHERME BUXBAUM MARINHO GUERRA

JOSE EDUARDO DUARTE MASSUCATO

LEONARDO CAETANO GOMIDE

LUCAS MESQUITA ANDRADE

VINICIUS LEITE CENSI FARIA

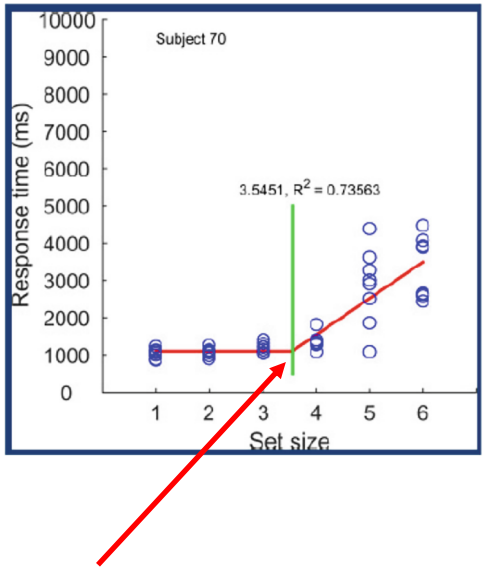
Padrões Minerados

- Recapitulando: o objetivo do artigo é identificar e compreender os padrões excepcionais relacionados às habilidades numéricas de crianças;
- Em específico, os autores investigam a tarefa "Dot Enumeration" (DE), pois ela engloba dois processos cognitivos importantes em crianças:
  - Subitizing
  - Counting
- O desempenho dessa tarefa é representada por uma função linear segmentada;
- Portanto, os autores, com base no desempenho das crianças em várias tarefas, tentam encontrar subgrupos onde a função do desempenho na tarefa DE, i.e. "subitizing curve", difere do da população geral;
- Formalmente, uma instância (criança) "r" é representada por

$$r = (a_1, \ldots, a_k, \ell_1, \ldots, \ell_m).$$

- mas no artigo, como para cada variável descritora podem ter tido mais do que uma entrada, r é alterado para:

$$r = ((a_{11}, a_{12}, ..., a_{1t}, ..., a_{1t_1}), (a_{21}, ..., a_{2t_2}), ..., (a_{k1}, ..., a_{kt_k}))$$



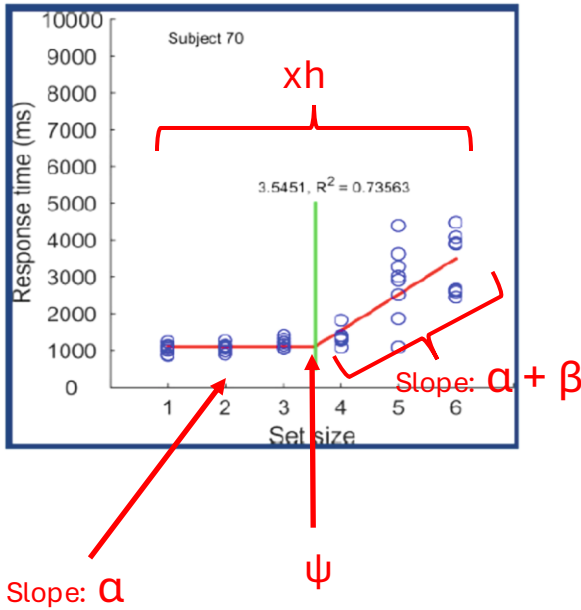
ponto de inflexão entre subitizing e counting

Regressão Linear Sementada: Target Model

- A regressão tem como objetivo descrever a relação entre o tempo de resposta (y) e o tamanho do conjunto (x_h). Ou seja, na tarefa de DE, dado um conjunto com N bolinhas, quanto tempo demora para a criança responder a quantidade especulada por ela.
- A fórmula é dada por:

$$y = g(x_h, \alpha, \beta, \psi) = \alpha x_h + \beta(x_h - \psi)_+$$

- Onde
  - α = inclinação do segmento de subitização
  - ψ = ponto de inflexão no x, que é o tamanho da faixa subitizing (esperado = 3 ou 4 pontos)
  - β = diferença entre a inclinação dos segmento subitizing e counting. A inclinação do segmento counting é então α + β
  - (xh − ψ)+ = função indicadora definida como (xh − ψ) · I(xh > ψ)
- Para facilitar:
  - Inclinação do primeiro segmento
    - Quando o tamanho do conjunto é menor que ψ **(xh ≤ ψ)**, o termo (xh − ψ)+ avalia para 0
    - Nesse intervalo, o modelo simplifica para **y = αxh**
  - Inclinação do segundo segmento
    - Quando o tamanho do conjunto é maior que ψ **(xh > ψ)**, o termo (xh − ψ)+ avalia para 1
    - Nesse intervalo, o modelo torna **y = αxh + β(xh − ψ)** = **(α + β)xh - βψ**



Descoberta de Subgrupos

		DE	
i	...	ℓ_1	ℓ_2
1	...	(5,1,8,...)	(1330,14...)
2	...	(8,2,1,...)	(2630,21...)
3	...	(4,9,2,...)	(2130,19...)
4	...	(7,4,5,...)	(2610,16...)

- O modelo, então, irá receber de entrada uma lista de observações acerca do desempenho de uma criança na tarefa DE
  - x_h = l1
  - y = l2
- **Porém**, tem um passo faltando, que é a obtenção dos subgrupos para que assim seja obtido as respectivas observações dos itens que os compõem. Para que isso ocorra, cada nova instância r deve ser modificada para que cada variável descritora tenha somente um valor. Assim, os autores agregam dados e extraem métricas de desempenho geral por tarefa.
- O subgrupo é definido agora por: $G_D = \{r^i \in \Omega | D(\tilde{a}_1^i, \tilde{a}_2^i, ..., \tilde{a}_s^i, a_{\dagger_1}^i, a_{\dagger_2}^i, ..., a_{\dagger_{k_\dagger}}^i) = 1\}.$
- E os novos atributos para cada instância r são:

Tasks	Name	Explanation
NC,SA,SS,CA	MaxItem	Number of answered items
NC,SA,SS,CA	SumAnsC	Number of correctly answered items
NC,SA,SS,CA	PropAnsC	Proportion of correctly answered items
NC,SA,SS,CA	MeanTC	Mean response time of correctly answered items
NC,SA,SS,CA	MedTC	Median response time of correctly answered items
NC,SA,SS,CA	IES	Inverse Efficiency Score
NC	IcNumD	Intercept of the response time regressed on the distance between the two numbers of every item
NC	SlNumD	Slope of the response time regressed on the distance between the two numbers of every item
NC	IcNumR	Intercept of the response time regressed on the ratio between the distance and the largest of the two numbers of every item
NC	SlNumR	Slope of the response time regressed on the ratio between the distance and the largest of the two numbers of every item

Medidas de Qualidade

- No processo da criação do modelo regressivo, queremos estimar o valor de ψ até que ele seja o valor mais provável. Tal valor é o maximum likelihood estimate.
- Os autores viram que maximizar o estimador é igual minimizar a sum-of-squares error (ssr). Portanto, os autores utilizam o **ssr** na **medida de qualidade do subgrupo.** Subgrupos com um ssr baixo e log likelihood alto indicam que o modelo ajustado se encaixa bem com os dados.
- Efetivamente, o artigo busca por subgrupos onde: $\ln p(SG|\theta^{SG}) > \ln p(SG|\theta^{\Omega})$
- Porém: $\ln p(y|\mathbf{x}, \mathbf{w}, \beta) \approx SSR(y, f(\mathbf{x}, \mathbf{w}))$
- Portanto, a primeira medida de qualidade é dada por:

$$\varphi_{\text{ssr}} = \frac{1}{\varphi_{\text{ef}}} \cdot -\frac{A}{N^{SG}} \quad A = SSR(\ell_2, g(\ell_1, \theta^{SG}))$$

- O problema com essa medida é que pode retornar um valor baixo para subgrupos onde a funcão de regressão é bem ajustada com os seus dados, mas que com os dados da populacão geral, também seria bem ajustada, ou seja, o subgrupo não é excepcional. Desta forma, os autores criaram a medida **ssrb**:

$$\varphi_{\text{ssrb}} = \varphi_{\text{ef}} \cdot \frac{A(B-A)}{N^{SG}}$$

- B é o A em **ssr**, mas com θ^{SG} substituído por θ^{Ω}

Metodologista

Bernnardo Seraphim

Diná Xavier

Gabriel Fadoul

Kênia Gonçalves

Oluwatoyin Joy

Samuel Kfuri

Metodologia

- Hipótese
- Características da base de dados
  - Flattening Approach
  - Funções Agregadas
- Exceptional Model Mining
  - Regressão Linear Segmentada
- Medida de Qualidade
  - Estimador de Máxima Verossimilhança
- Experimentos
- Conclusões

Hipótese

- Como identificar de maneira robusta grupos de crianças que apresentam padrões de subitização atípicos?
- A abordagem de Mineração de Modelos Excepcionais pode descobrir relações mais complexas nas variáveis de interesse.
- Estratégia Metodológica:
  - Pré-Processamento: transformação de base de dados complexos em um formato padrão
  - Mineração de Modelos: uso do algoritmo de busca para encontrar subgrupos com "modelos" de comportamento excepcionais
  - Validação: Definir e utilizar métricas de qualidade para avaliar e validar os subgrupos encontrados

Características da Base de Dados

- FUnctional Numerical Assesment (FUNA)
  - Estudo em larga escala com o intuito de detectar discalculia e dislexia em estudantes da Finlândia e Suécia
- A base de dados não apresenta um formato tradicional
  - Tarefas realizadas de maneira cronometrada
  - Tarefas realizadas em ordem quasi-aleatória (mais fáceis primeiro)
  - Múltiplas medições para vários itens
- Solução: "Achatamento" por Agregação de Atributos
  - E.g. Tempo médio de resposta dos itens corretos, proporção de respostas corretas, métricas observáveis obtidas da bibliografia (IES – Inverse Efficiency Score)

EMM – Exceptional Model Mining

- Uso de busca em feixe associado ao esquema de cobertura ponderada (***WCS -*** *Weighted Coverage Scheme*)
  - Diminui o "peso" de instâncias que já foram cobertas por subgrupos encontrados
  - Previne repetição de resultados
- O modelo escolhido é a regressão linear por partes (Segmented Linear Regression) visto que é um bom modelo para representar o diferente comportamento entre a subitização e a enumeração

Medidas de Qualidade

- Foram apresentadas duas medidas de qualidade
  - **SSR** → Leva em consideração o erro do modelo do subgrupo
  - **SSRB** → Leva em consideração o erro do modelo do subgrupo em relação ao erro do modelo global

Experimentos

- Experimento 1
  - Feito sobre 5% do data set
  - Validação das métricas de qualidade propostas e ajuste de parâmetros como profundidade de busca (d) e o parâmetro γ do WCS
  - Qualidade média, tamanho dos subgrupos, redundância e tempo de execução são métricas avaliadas
- Experimento 2
  - Extensivo sobre todo o dataset
  - Objetivo de descobrir quais são os subgrupos excepcionais e o que eles dizem
  - Validação se características de subgrupos condiziam com o padrão de subitização encontrado

Resultados Experimento 1

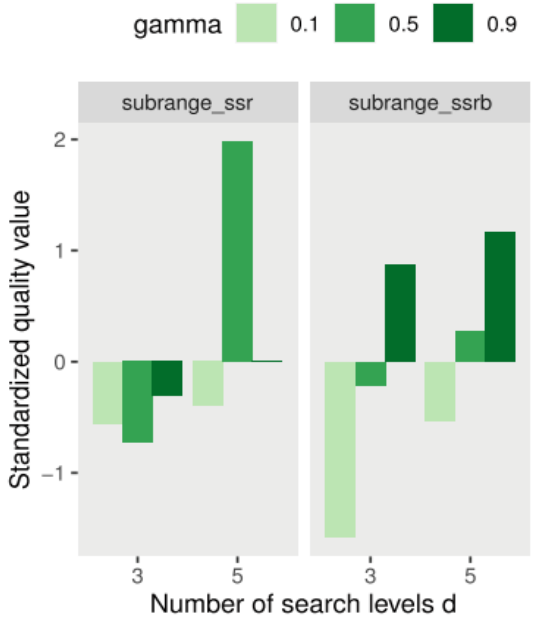


Fig. 3. The relation between the average quality of a subgroup set ($q = 10$, standardized per QM), search depth d, and WCS parameter γ, for both QMs.

QM	d	γ	Prop	DFD	JE	JSIM	Time
φ_{ssr}	3	0.1	0.20	0	1.36	0.87	1.63
		0.5	0.16	0	1.68	0.75	1.61
		0.9	0.16	0	1.55	0.73	1.62
	5	0.1	0.12	0	0.91	0.88	2.97
		0.5	0.13	0	0.79	0.91	2.95
		0.9	0.13	0	0.77	0.90	2.93
φ_{ssrb}	3	0.1	0.22	2	4.35	0.18	1.60
		0.5	0.08	0	2.35	0.31	1.56
		0.9	0.05	0	1.14	0.44	1.35
	5	0.1	0.06	0	2.19	0.29	2.17
		0.5	0.06	0	2.01	0.31	2.13
		0.9	0.05	0	1.08	0.46	1.86

Fig. 4. Experimental results for both QMs, $d \in \{3, 5\}$, $\gamma \in \{0.1, 0.5, 0.9\}$.

- **SSRB**(que compara o erro do subgrupo com o erro do modelo global) foi escolhida para o experimento principal por ser mais estável e produzir subgrupos menores e mais interessantes.
- Uma profundidade de busca de **d=3** foi selecionada por não apresentar grande diferença em relação a **d=5** e por gerar descrições mais fáceis de interpretar.
- Um valor de **γ=0.5** foi escolhido para equilibrar a busca por alta qualidade com a necessidade de baixa redundância nos resultados.

Resultados Experimento 2

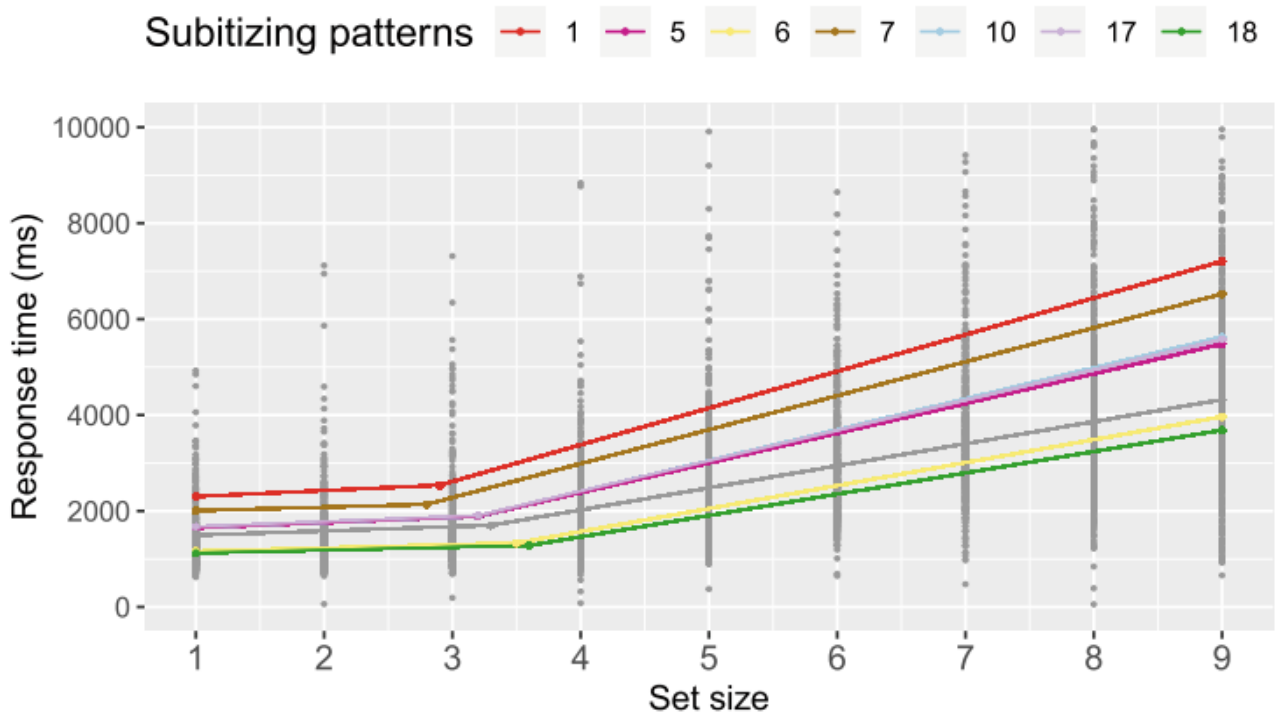


Fig. 5. Estimated segmented linear regression models of subgroups 1, 5, 6, 7, 10, 17 and 18 discovered with φ_{ssrb}. Target model equations can be found in Table 3.

Resultados Experimento 2

Table 3. Subgroup proportion, description and estimated target models for subgroups 1, 5, 6, 7, 10, 17 and 18, discovered with φ_{ssrb}. The global target model is $1407 + 88\ell_1 + 463(\ell_1 - 3.3)_+$

SG	Prop	Description	Target model
1	0.05	NC-IES:(0.04,1.0) $\wedge$ NC-MeanTC:(0.23,0.74) $\wedge$ SA-MeanTC:(0.71,1.0)	$2179 + 124\ell_1 + 764(\ell_1 - 2.9)_+$
5	0.50	NC-IES:(0.03,1.0)	$1541 + 106\ell_1 + 624(\ell_1 - 3.2)_+$
6	0.50	NC-IES:(0,0.03)	$1091 + 70\ell_1 + 475(\ell_1 - 3.5)_+$
7	0.12	NC-MeanTC:(0.16,1.0) $\wedge$ SS-SumAnsC:(0.0,0.32)	$1938 + 70\ell_1 + 712(\ell_1 - 2.8)_+$
10	0.38	SA-MeanTC:(0.71,1.0)	$1544 + 108\ell_1 + 641(\ell_1 - 3.2)_+$
17	0.30	grade $= 3$	$1561 + 106\ell_1 + 634(\ell_1 - 3.2)_+$
18	0.27	SA-MaxItem:(0.6,1.0)	$1064 + 62\ell_1 + 441(\ell_1 - 3.6)_+$

Conclusão

- Pontos Fortes
  - A metodologia de agregação de dados é uma solução elegante dado a complexidade do dataset
  - A escolha do modelo de regressão por partes é justificada pela natureza do problema
  - O uso de técnicas como WCS torna a metodologia robusta e os resultados mais confiáveis
- Limitações
  - O modelo de regressão assume que múltiplas observações de uma mesma criança são independentes
  - O modelo pressupões que o ponto de quebra entre subitização e enumeração é pontual para o estudante, o que não é necessariamente

Assessor Social

Caio Jorge Carvalho Lara
Juan Marcos Braga Faria
Luisa Vasconcelos de Castro Toledo
Luiza Sodré Salgado
Mateus Reis Evangelista
Samuel Henrique Miranda Alves

Pontos positivos

- O artigo trabalha com uma técnica avançada de mineração de dados (EMM)
  - Captura correlação entre 2 variáveis, permitindo a modelagem de eventos mais complexos do mundo real
  - Dificulta a interpretabilidade, mas encontra resultados mais profundos

- Essas técnicas podem ser aplicadas para descrever padrões inicialmente não perceptíveis em uma base de dados

Pontos positivos

- Identificação de deficiências na fase infantil pode ser crucial
  - Eficiência no tratamento
- Aprimoramento de técnicas pedagógicas com crianças
- Pode ser estendido para outras áreas do conhecimento
  - Pedagogia: extensão da pesquisa para disciplinas de linguagens ou outras faixas etárias
  - Psicologia: identificar comportamentos específicos em transtornos

Pontos negativos

- Modelos computacionais são "limitados"
  - Aprendem apenas com os dados que são repassados a eles
  - Podem existir variáveis de confusão no meio
  - Dificuldades ocasionais em matemática, por exemplo, nem sempre significam um quadro de discalculia
- Risco de privacidade
  - Informações sobre a saúde são dados sensíveis
  - Identificação de crianças com quadro de discalculia pelo cruzamento de dados externos

Pontos negativos

- Risco de estigmatização
  - Identificar subgrupos excepcionais pode criar rótulos negativos
- Equidade algorítmica
  - Vieses podem surgir da falta de qualidade dos dados
  - Crianças com menor familiaridade digital podem ser indevidamente classificadas como atípicas
- Necessidade de transparência da interpretação dos resultados

Pontos negativos

- Psicologia infantil é um tema complexo
  - Crianças podem apresentar diferentes evoluções de aprendizado
  - Dificuldades de aprendizado podem, por exemplo, não ser ligados a transtornos, mas a variáveis externas (ambiente familiar)

- Apesar do artigo aplicar uma técnica avançada, ainda assim falham em capturar todas as variações
  - Existe uma forte presunção de que os modelos dos subgrupos seguem o mesmo padrão da tendência global

- Limitação dos dados
  - Os dados são apenas de crianças finlandesas (não é generalizável)

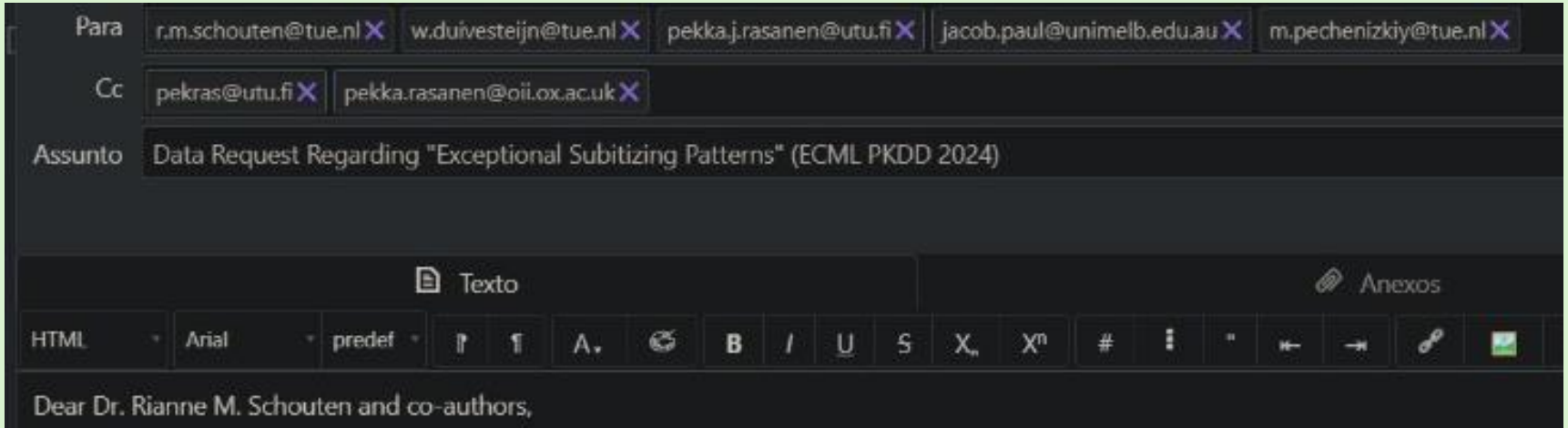
Alexis Duarte Guimaraes Mariz
Amanda Mendes Pinho
Gabriel Chaves Ferreira
João Vítor Fernandes Dias
Kael Soares Augusto
Lucas Xavier Veneroso

Busca pelo código

- 6. Experiments: https://github.com/RianneSchouten/FUNA_EMM

Busca pelos dados

- Dados disponíveis:
  - Curran: 100%
    - sav: data (405x15); long (1393x14)
    - pq: desc (405x17); long (1325x8); target (1325x4); wide (405x18)
  - FUNA: 0,05%
    - pq: Target (13135x4); Desc (774x32)
- "More data can be made available upon request"

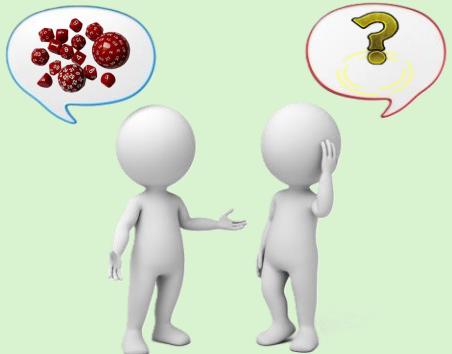


Busca pelos dados

- Dados solicitados: ~~SPAM~~ Solicitação com redundância.
- Pekka Räsänen
  - […] representative of the owner of the datasets and the test, […] your plans to use the data before we can consider giving it for your work. The data is not open-access, and not available to master's, or below, level work. Please send me an official request including the names and contact information of your supervisors and a detailed research plan.
- Outros caminhos: Projeto FUNA; Issue; LinkedIn

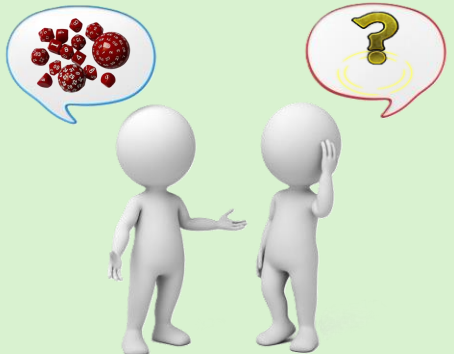
Entendendo os dados encontrados (FUNA)

- Desc
  - float64: 28 colunas
    - {NC, SA, SS, CA} x {AnsCsum, PreOrdmax, AnsCprop, timeCmean, timeCmedian, IES}
    - {NCRT} x {slopeNumDis, interceptNumDis, slopeNumRatio}
  - category: 1 coluna {grade}
  - string: 3 colunas {IDCode, sex, language}
- Target
  - int64: 3 colunas {DMStimL, DMTime, PreOrd}
  - string: 1 coluna {IDCode}



Entendendo os dados encontrados (Curran)

- SAV: Statistical Package for the Social Sciences
- Data
  - float64: 14 colunas
    - {anti, read} x {1, 2, 3, 4}
    - {homecog, homeemo, id, kidage, momage, nmis}
  - category: 1 coluna {kidgen}
- Long
  - float64: 13 colunas
    - {kidage} \times {Ø, 6, c, sq, tv}
    - {anti, homecog, homeemo, id, momage, occasion, occasion2, read}
  - category: 1 coluna {kidgen}



Preparando o código

- "The virtual environment can be start with `venv\Scripts\activate"`

.gitignore

```
123
124    # Environments
125    .env
126    .venv
127    env/
128    venv/
129    ENV/
130    env.bak/
131    venv.bak/
```

- Correção do `requirements.txt` com versões compatíveis
  - `python -m venv venv`
  - `venv/Scripts/activate`
  - `pip install –r requirements.txt`
- Testes: main.py: 9 casos de teste comentados

Entendendo os testes: Parâmetros

- Básicos
  - datasets_names=['desc']
  - synthetic_params=None: Nunca ocorre, a função inclusive nem existe
  - date: dia da execução
  - data_name: 'FUNA' ou 'Curran'
  - data_from: input path
  - output_to: output path

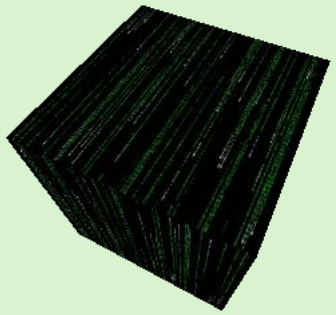
Entendendo os testes: Parâmetros

- sim_params
  - b=[4]: quantidade de quantis
  - w=[20]: critério de parada do dbs
  - d: profundidade da hierarquia
  - q: quantidade de grupos
  - model: modelos de estimativas
  - dbs=[False]: Se True (nunca é), aplica description-based selection ao preparar o Beam
  - wcs=[True]: Se true (sempre é), aplica cover-based selection ao preparar o Beam
  - gamma: vetor de pesos
  - dp=[False]: Se True (nunca é), aplica dominance pruning no Beam Search
  - md=['without']: influences selection in ss
  - min_size=0.05: tamanho mínimo do subgrupo

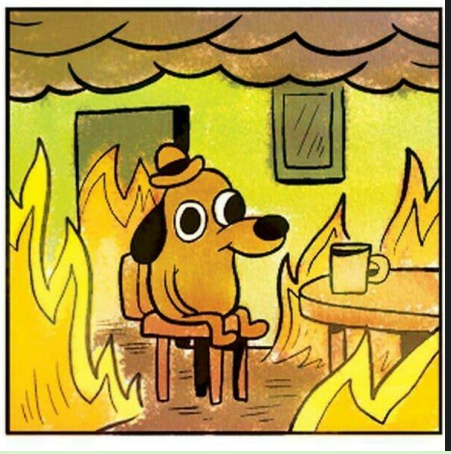
Entendendo os testes: Parâmetros

- extra_info
  - target_column_names: lista de nomes das colunas destino. Sempre os mesmos pra FUNA e Curran
  - sample: determina se será feita uma amostragem
  - sample_prop: proporção da amostragem
  - case_based_target=False: if True only select first available row of every case
  - run_redun_metrics: executa métricas de redundância
  - run_beam_search=True: executa busca em feixe
  - make_dfd: Se True, cria uma Distribuição de Falsas Descobertas
  - m: range para percorrer a distribuição de falsos descobertas

Rodando o código



- Curran: 1h30min
- E o FUNA?



```
(venv) PS B:\GitHub\UFMG\FUNA_EMM> python main.py
importing data
Data format 1 out of 1 : desc
Simulation 1 out of 18 : (4, 20, 3, 10, 'reg_ssr', False, 0.05, True, 0.1, False,
'without', 0.05)
d i 1
|                                                  | 92/92 [100%] in 0.9s (105.48/s)
d i 2
|                                                  | 20/20 [100%] in 25.4s (0.77/s)
d i 3
|                                                  | 20/20 [100%] in 21.5s (0.91/s)
|
```

ult.py - FUNA_EMM - Visual Studio Code

PROBLEMAS 1K+ SAÍDA CONSOLE DE DEPURAÇÃO

- beam_search.py beam_search 44
- measures.py beam_search 90
- parameters.py beam_search 56
- prepare_beam.py beam_search 34
- qualities.py beam_search 104
- refinements_functions.py beam_search 28
- refinements.py beam_search 46
- select_subgroup.py beam_search 78
- summarize.py beam_search 7
- constraints.py constraints 14
- cover_based_selection.py constraints 24
- desc_based_selection.py constraints 28
- dominance_pruning.py constraints 24
- analysis.py experiment 51
- distribution_false_discoveries.py experiment 35
- process_result.py experiment 74
- pwlf.py experiment 92
- retrieve_curran.py experiment 36
- retrieve_funa.py experiment 58
- retrieve_rw_data.py experiment 20
- save_and_store_result.py experiment 38
- main.py 46

Erros: 692, Avisos: 82, Informações: 268

692 82 268 Live Share

Outputs

- Disponibilizados na pasta output
- Cada pasta tem uma lista de arquivos txt com a descrição dos parâmetros usados
- E também um arquivo xlsx com os parâmetros e as métricas encontradas

Referências

Referências bibliográficas

- SCHOUTEN, R. M. et al. Exceptional subitizing patterns: Exploring mathematical abilities of Finnish primary school children with piecewise linear regression. Em: Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2024. p. 66–82.