Avaliação do Impacto de Estratégias de Pré-processamento de Sequências de Eventos de Aprendizagem em Algoritmos de Mineração de Padrões Sequenciais

José Thiago Torres da Silva, Júlio César Roque da Silva, Patricia Takako Endo, Raphael A. Dourado

[1] Universidade de Pernambuco (UPE), campus Caruaru, 55.014-908 – Caruaru, PE

{thiago.torress,julio.roque,patricia.endo,raphael.dourado}@upe.br

Resumo. *Dados relativos a eventos de aprendizagem, quando contêm atributos temporais, permitem analisar a aprendizagem de fato como um processo ao longo do tempo utilizando, por exemplo, algoritmos de Mineração de Padrões Sequenciais (Sequential Pattern Mining - SPM). No entanto, são escassos na literatura atual trabalhos que avaliam o impacto de estratégias de pré-processamento destas sequências de eventos nos padrões identificados pelos algoritmos. Este estudo investiga o impacto de três estratégias de pré-processamento propostas na literatura nos padrões identificados pelo algoritmo PrefixSpan, utilizando uma base de dados real de cursos à distância oferecidos na plataforma Moodle. Os resultados foram analisados de forma quantitativa e qualitativa e sugerem que a estratégia "Coalescing Repeating Point Events into One" teve o maior impacto na remoção de ruídos, embora o uso conjunto das três estratégias contribuiu para melhorar a qualidade dos padrões detectados.*

Abstract. *Timestamped learning event datasets make it possible to analyze learning in fact as a process over time using, for example, Sequential Pattern Mining (SPM) algorithms. However, there are few studies that evaluate the impact of preprocessing strategies for event sequences on the patterns identified by the algorithms. This study investigates the impact of three preprocessing strategies proposed in the literature on the patterns identified by the PrefixSpan algorithm, using a real database of distance learning courses offered on the Moodle platform. We analyzed the results both quantitatively and qualitatively. The findings suggest that the "Coalescing Repeating Point Events into One" strategy had the greatest impact on noise removal, although the joint use of the three strategies contributed to improve the quality of the detected patterns.*

1. Introdução

O uso de Ambientes Virtuais de Aprendizagem (AVAs) como ferramentas de auxílio no ensino presencial, híbrido e educação à distância (EaD) tem possibilitado a coleta e armazenamento de uma grande quantidade de dados relativos às interações dos estudantes, professores e outros autores com estes ambientes [Peña-Ayala 2023, Wise 2019]. Esta crescente disponibilidade de dados relativos a atividades de ensino-aprendizagem junto à evolução das técnicas de Mineração de Dados fomentou o surgimento da Mineração de Dados Educacionais, ou EDM (do inglês *Educational Data Mining*), que busca aplicar técnicas de mineração de dados adaptadas às especificidades do contexto educacional.

DOI: 10.5753/sbie.2024.242734

Estas bases de dados educacionais podem ser analisadas com diversos objetivos e fins pedagógicos. Wise [Wise 2019] lista como principais objetivos a predição, descoberta de estruturas (como correlação e regras de associação), processamento de linguagem natural, visualização de dados e abordagens temporais. Nesta última categoria, o foco é em descobrir padrões onde o tempo é um componente central, viabilizando assim analisar a aprendizagem de fato como um processo ao longo do tempo. Outros autores classificam esta abordagem como "process analytics" [Lockyer et al. 2013].

Análises deste tipo podem ser viabilizadas utilizando técnicas e algoritmos consolidados de Mineração de Padrões Sequenciais (*Sequential Pattern Mining* - SPM), já que as trajetórias de aprendizagem de estudantes podem ser modeladas computacionalmente como sequências de eventos [Zhang and Paquette 2023]. No entanto, além da carência de trabalhos que de fato analisam os processos educacionais levando em conta os aspectos temporais [Zhang and Paquette 2023, Verbert et al. 2020, Wise and Jung 2019, Bogarín et al. 2018], há poucas evidências na literatura sobre o impacto de estratégias de pré-processamento de sequências de eventos nos padrões de aprendizagem descobertos pelos algoritmos de SPM.

Assim, este trabalho tem como objetivo investigar o impacto de três estratégias de pré-processamento de sequências de eventos propostas na literatura em uma base de dados real de logs de interação de alunos com um curso à distância na plataforma Moodle. O restante do texto está estruturado da seguinte forma: a Seção 2 detalha os principais conceitos teóricos utilizados no trabalho e discute os trabalhos relacionados; a Seção 3 apresenta o método utilizado; a Seção 4 descreve e discute os resultados obtidos; e a Seção 5 delineia as principais conclusões e limitações e aponta possibilidades de trabalhos futuros.

2. Fundamentação Teórica e Trabalhos Relacionados

Nesta seção, são apresentados os conceitos de Mineração de Dados Educacionais, análise processual da aprendizagem, sequências de eventos, mineração de sequências e estratégias de pré-processamento. São discutidos também trabalhos relacionados e suas diferenças com o presente estudo.

2.1. Mineração de Dados Educacionais e Análise Processual da Aprendizagem

A Mineração de Dados Educacionais (em inglês, *Educational Data Mining* - EDM) preocupa-se principalmente em desenvolver métodos para explorar as características únicas dos dados coletados em ambiente educacionais e utilizar estes métodos para entender o comportamento dos estudantes e as circunstâncias em que ocorre a aprendizagem [Peña-Ayala 2023, Chen et al. 2020, Baker 2014]. A EDM surge ao mesmo tempo como uma ciência do aprendizado e também como uma rica área de aplicação para Mineração de Dados, dado a crescente disponibilidade de dados relativos a contextos educacionais. Assim, a EDM possibilita decisões baseadas em evidências para aperfeiçoar as práticas e materiais de ensino [Calders and Pechenizkiy 2012].

Para além das linhas de pesquisa mais comuns em EDM focadas em tarefas como predição de desempenho/evasão e identificação de agrupamentos (clusters), um tipo de análise que tem recebido menor atenção é a análise processual da aprendizagem, definida por Lockyer et al. [Lockyer et al. 2013] como "*process analytics*". Neste tipo de análise,

que Wise [Wise 2019] chama de "aspectos temporais da aprendizagem", o tempo é um componente central, priorizando fatores como a ordem e o momento em que os estudantes executam tarefas ao invés de apenas quantificadores como número de acessos ao ambiente e entregas de atividades. Neste trabalho, o componente temporal *ordem de acesso a recursos educacionais* é o foco de análise, o que justifica a escolha de técnicas de análise de sequências de eventos, discutidas a seguir.

2.2. Sequências de Eventos e *Sequential Pattern Mining* (SPM)

Uma sequência de eventos pode ser definida como uma lista ordenada de eventos $S = [e_0, e_1, \ldots, e_n]$, onde cada elemento $e = (\tau, t)$ representa um evento distinto, sendo τ o tipo do evento e t o momento (*timestamp*) em que o evento ocorreu [Guo et al. 2022]. Todos os eventos e_n devem pertencer a um conjunto de símbolos $I = \{i_1, i_2, ..., i_m\}$, que representa o domínio dos tipos de eventos possíveis na base de dados sob análise.

Este tipo de abstração pode ser utilizada para modelar computacionalmente sequências de eventos de aprendizagem em AVAs [Zhang and Paquette 2023, Guo et al. 2022]. Neste contexto, o conjunto I representa os tipos possíveis de eventos de aprendizagem registrados pelo sistema de log do AVA, e S os eventos da trajetória de aprendizagem de um estudante.

Uma das tarefas possíveis com sequências de eventos é a Mineração de Padrões Sequenciais (ou em inglês, *Sequential Pattern Mining* - SPM[1]). Originalmente proposta por Agrawal & Srikant [Agrawal and Srikant 1995], a SPM tem como objetivo encontrar subsequências recorrentes em bases de dados de sequências. Por exemplo, considerando uma base com duas sequências, $S_1 = [A, B, C, D]$ e $S_2 = [A, B, D, C]$, o padrão A, B é comum a ambas. Alguns parâmetros de restrição (*constraint*) são comumente utilizados para reduzir o espaço de busca dos algoritmos de SPM. Um deles é o "*gap*" [Fournier Viger et al. 2017], que define uma distância máxima entre os eventos ao considerá-los como um padrão, sendo o valor padrão 1. Ainda utilizando as sequências exemplo S_1 e S_2, com gap = 1 apenas o padrão A, B seria detectado, enquanto com gap = 2 o padrão B, C também seria considerado, pois ocorre em S_2 com apenas um evento intermediário (D). Outros parâmetros comumente utilizados são o tamanho mínimo e máximo dos padrões a serem buscados.

A avaliação da saída dos algoritmos de SPM é normalmente realizada utilizando a métrica de *suporte* [Fournier Viger et al. 2017]. O suporte é calculado para cada padrão (subsequência) identificado pelo algoritmo como a quantidade de sequências da base onde aquele padrão foi encontrado, podendo ser expresso em números absolutos ou percentuais. Outras métricas que podem ser utilizadas são o tamanho dos padrões (quantidade de eventos) e *I-support* [Lo et al. 2008, Zhang and Paquette 2023]. Estas três métricas foram utilizadas neste estudo e serão explicadas em mais detalhes na Subseção 3.4.

Por fim, existem diversos algoritmos propostos na literatura para mineração de padrões sequenciais. No entanto, por ser uma tarefa de resultado determinístico, qualquer algoritmo, dados os mesmos parâmetros, retornará o mesmo resultado, de forma que a escolha em geral é feita por questões de performance e parâmetros suportados. Alguns dos algoritmos mais populares são o GSP (Generalized Sequential Pattern)

[1]Por já ser um acrônimo consolidado, será utilizada a sigla SPM ao longo do trabalho.

[Agrawal and Srikant 1995], cSPADE [Zaki 2000] e PrefixSpan [Pei et al. 2001]. O PrefixSpan, utilizado neste trabalho, tem como principais parâmetros a faixa de tamanho dos padrões a serem identificados e o valor de suporte mínimo desejado. Ele funciona em 3 passos: 1) busca por padrões sequenciais de tamanho 1, chamados de prefixos, que são utilizadas para buscar as próximas subsequências; 2) divisão do espaço de busca, dividido em subconjuntos de acordo com as sequências que começam com cada um dos prefixos identificados anteriormente; e 3) encontrar subsequências identificando padrões, fazendo uma mineração recursiva pelos subconjuntos.

2.3. Pré-processamento de Sequências de Eventos

Conforme defendido por Zhang & Paquette [Zhang and Paquette 2023], a aplicação de algoritmos de SPM em dados educacionais requer uma etapa de pré-processamento das sequências de forma a evitar que os algoritmos detectem padrões inconsistentes ou irrelevantes. Esta etapa é especialmente importante em dados educacionais devido ao alto nível de ruído que os sistemas de log dos AVAs, por exemplo, podem introduzir.

Existem na literatura propostas de estratégias de pré-processamento de dados específicas para sequências de eventos. Du et al. [Du et al. 2017] propõem um conjunto de estratégias para pré-processamento de sequências de eventos já aplicadas com sucesso em outros estudos na literatura. Han e Kamber [Han and Kamber 2012] também compilam estratégias para redução e enriquecimento de bases de sequências de eventos. Neste trabalho, foram utilizadas cinco destas estratégias, escolhidas com base em sua adequabilidade para a natureza dos dados a serem analisados. São elas:

- *Temporal Windowing* - Em determinados casos, não é necessário analisar toda a extensão temporal dos dados disponíveis, mas apenas uma determinada janela de tempo que contém as informações mais relevantes. Esta estratégia propõe a definição de uma janela de tempo para análise, desconsiderando o restante dos dados [Du et al. 2017].
- *Goal-Driven Event Category Extraction* - Bases de dados de sequências de eventos podem apresentar um conjunto grande de categorias de eventos, o que pode introduzir ruídos que atrapalham o desempenho dos algoritmos de SPM. Em muitos casos, nem todas essas informações são relevantes para a análise. Essa estratégia, geralmente utilizada no começo do processo de extração de dados, propõe que sejam removidas as categorias de eventos que não são essenciais para o objetivo da análise [Du et al. 2017].
- *Multilevel Sequential Patterns* - Eventos podem conter, além de sua categoria, outras informações associadas, as quais, segundo Han & Kamber [Han and Kamber 2012], podem ser utilizadas para enriquecer as sequências e revelar padrões mais relevantes. Isso se dá através da adição desses "*levels* (níveis)" às categorias originais de eventos (criando assim novas categorias) que serão levadas em consideração pelos algoritmos durante a busca de padrões.
- *Coalescing Repeating Point Events into One* - Existem alguns tipos de eventos que, acontecendo repetidamente de forma sequencial, acabam não trazendo nenhuma informação nova sobre a sequência. Nestes casos, Du et al. [Du et al. 2017] sugere que apenas a primeira ou última ocorrência seja mantida.
- *Converting Hidden Complex Events into One* - Subsequências de eventos podem ocorrer repetidamente em uma sequência e, ao invés de representar um padrão de

interesse, ser apenas um evento complexo registrado em baixa granularidade pelo ambiente que coletou os dados. Um exemplo comum são caminhos em interfaces onde uma ação depende obrigatoriamente de outra, como o acesso a um carrinho de compras antes de remover um item. Du et al. [Du et al. 2017] argumenta que estes eventos complexos podem ser simplificados para um evento único, reduzindo o tamanho das sequências e evitando a detecção de padrões óbvios.

Os detalhes de como estas estratégias foram utilizadas com os dados deste estudo serão descritos na Seção 3.

2.4. Trabalhos Relacionados

Song et al [Song et al. 2022], utilizando dados de MOOCs (*Massive Open Online Courses*), propõem uma nova métrica de suporte (*support with flexible constraints - SFC*) para seleção de padrões frequentes em sequências de eventos baseada no uso de restrições flexíveis (*flexible constraints*). Os autores propõem três restrições: tamanho da sequência de eventos, variância temporal dos eventos dentro de uma sequência (discreteness) e validade, que no contexto estudado pelos autores distingue eventos de acesso casual e "sérios" (casual/serious learning). No entanto, o foco do trabalho não é no impacto de estratégias de pré-processamento sobre as sequências resultantes, mas sim na redução do espaço de busca de algoritmos de SPM e, consequentemente, no aumento da performance dos mesmos.

Poon et al. [Poon et al. 2017] utilizaram em seu estudo duas estratégias do que [Zhou et al. 2010] chamam de "modelagem de sequências": i) divisão dos estudantes em grupos com base no seu desempenho (nota) e ii) filtragem de estudantes com base em recursos educacionais de interesse (e.g. incluir apenas estudantes que assistiram aos vídeos disponibilizados). Os autores analisam o impacto do uso destas duas estratégias no tempo de execução do algoritmo de SPM e no número de sequências identificadas pra diferentes níveis de suporte mínimo (entre 0,2 e 1). No entanto, outras características como suporte, i-support, tamanho das sequências e relação com o desempenho dos alunos não são analisadas.

Munk *et al.* [Munk et al. 2017] avaliaram o impacto de diferentes estratégias de identificação de sessões e finalização de caminhos (*path completion*) em mineração de sequências de eventos, enquanto Maranhão *et al.* [Maranhão et al. 2023] demonstraram como SPM pode ser utilizada para identificar possíveis problemas que os(as) estudantes enfrentam em exercícios de programação e que possam ser causa de desmotivação aos mesmos. No entanto, ambos os trabalhos não avaliam diferentes estratégias de pré-processamento, como proposto no presente estudo.

Desse modo, o diferencial deste trabalho em relação ao estado da arte é a avaliação do impacto de diferentes técnicas de pré-processamento de sequências de eventos na quantidade e qualidade dos padrões identificados por algoritmos de SPM utilizando métricas propostas na literatura.

3. Método

Esse estudo foi desenvolvido seguindo as etapas propostas pela metodologia CRISP-DM (*CRoss Industry Standart Process for Data Mining*) [Azevedo and Santos 2008,

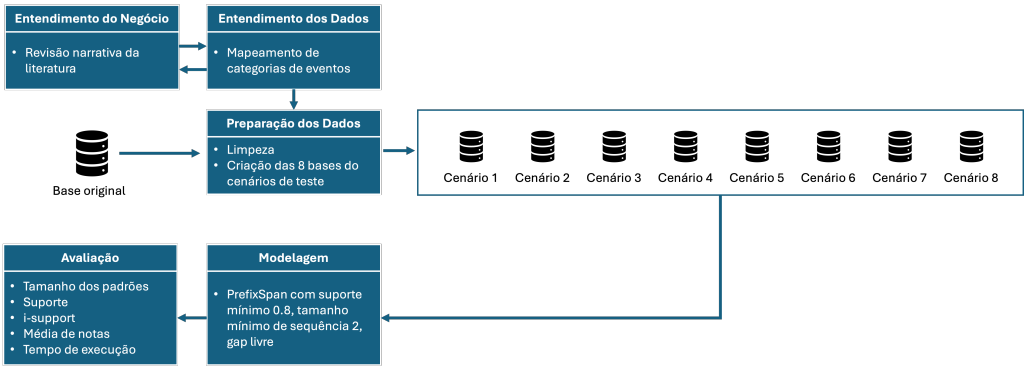

Figura 1. Fases do método utilizado no estudo, seguindo a metodologia CRISP-DM

Schröer et al. 2021], largamente utilizada em projetos de Mineração de Dados, que propõe seis etapas iterativas: entendimento de negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação. A instanciação destas fases, com exceção da fase de implantação, que não se aplica aos objetivos deste estudo, é ilustrada na Figura 1 e descrita a seguir.

3.1. Entendimento do negócio e Entendimento dos dados

Para entendimento do negócio, foi realizada uma revisão narrativa da literatura sobre EDM, AVAs, *process analytics* [Lockyer et al. 2013] e análise de sequências de eventos aplicada a dados educacionais [Zhang and Paquette 2023]. Já na etapa de entendimento dos dados, foi realizado o estudo da base de dados utilizada neste artigo, composta por dados reais de uma instância da plataforma Moodle, disponibilizada para a equipe do projeto por uma instituição de ensino pública que oferece cursos técnicos na modalidade de EaD. A base foi devidamente anonimizada pela instituição de ensino antes de ser entregue à equipe de pesquisa, removendo todas as informações sensíveis de alunos e professores.

Para extração das sequências de eventos de aprendizagem dos estudantes, foi utilizada a tabela de logs do Moodle (mdl_logstore_standard_log), que registra ações dos usuários com carimbo de tempo (*timestamp*). Estas ações foram mapeadas a dez classes de eventos de aprendizagem com base nos atributos “component”, “action” e “target” da tabela mdl_logstore_standard_log, conforme descrito na Tabela 1.

Para este estudo, foram utilizados os registros de interação dos alunos do Curso Técnico em Gestão de Recursos Humanos durante o período de realização da primeira atividade da disciplina “Cargos e Salários”, um quiz composto por quatro perguntas de múltipla escolha. A fim de analisar todo o percurso de aprendizagem dos alunos, foi considerado o período desde a publicação da atividade no sistema até a data de entrega. A Tabela 2 fornece uma visão geral desta base de dados, evidenciando uma grande heterogeneidade no tamanho das sequências de aprendizagem de cada aluno.

3.2. Preparação dos dados

A preparação dos dados foi iniciada com uma limpeza das sequências de eventos de cada aluno para remover redundâncias e eventos não relacionados ao desenvolvimento da atividade sob análise. Para isso, foram utilizadas duas estratégias básicas de preparação de sequências propostas por [Du et al. 2017]:

Tabela 1. Mapeamento dos logs registrados pelo Moodle a classes de eventos de aprendizagem

component	**action**	**target**	**Classe**	Descrição
core	viewed	course	course_vis	Visualização da página da disciplina
mod_page	viewed	course_module	course_vis	
mod_chat	sent	message	message_sent	Mensagem privada enviada
mod_chat	viewed	course_module	message_read	Mensagem privada lida
mod_folder	downloaded	all_files	resource_vis	Visualização dos recursos
mod_folder	viewed	course_module	resource_vis	
mod_resource	viewed	course_module	resource_vis	
mod_url	viewed	course_module	resource_vis	
mod_forum	created	discussion	forum_participation	Participação no fórum
mod_forum	created	post	forum_participation	
mod_forum	updated	post	forum_participation	
mod_forum	created	subscription	forum_followup	Inscrição para acompanhar um tópico do fórum
mod_forum	viewed	course_module	forum_vis	Visualização do fórum
mod_forum	viewed	discussion	forum_vis	
mod_quiz	started	attempt	assignment_try	Tentativa de realizar a atividade
mod_quiz	submitted	attempt	assignment_sub	Entrega da atividade
mod_quiz	viewed	course_module	assignment_vis	Visualização da atividade

Tabela 2. Características do dataset utilizado

Característica	Quantitativo
Número de alunos/sequências	2085
Tamanho da maior sequência	127
Tamanho da menor sequência	3
Média de tamanho das sequências	11,34
Desvio padrão do tamanho das sequências	12,02

- *Temporal Windowing* - Foram removidos todos os eventos ocorridos após a última submissão da atividade (considerando que o aluno pode submeter mais de uma vez).
- *Goal-Driven Event Category Extraction* - O evento relativo à visualização do curso (course_vis) é um dos mais comuns na base, porém não possui carga semântica relevante para a análise das sequências de aprendizagem (é pré-requisito o aluno acessar o curso para realizar qualquer outra ação). Assim, foram removidas todas as ocorrências deste evento nas sequências.

Em seguida, foram escolhidas três técnicas propostas na literatura por Du et al. [Du et al. 2017] e Han & Kamber [Han and Kamber 2012] para avaliação neste estudo, já apresentadas na Seção 2, as quais foram aplicadas à base de dados da seguinte forma:

- *Multilevel Sequential Patterns* - Considerando a importância do quão cedo ou tarde o aluno realizou uma ação em relação à data limite para a entrega da atividade, foi adicionado um sufixo aos eventos para representar este dado, adicionando assim um novo nível aos eventos de aprendizagem. Em todos os eventos que ocorreram até 50% do tempo do prazo da atividade foi adicionado o sufixo *_START*, e aos demais o sufixo *_END*.
- *Coalescing Repeating Point Events into One* - Na base original, é comum que os alunos realizem um mesmo evento repetidas vezes, como por exemplo,

visualização de vários tópicos de discussão em um mesmo fórum num curto espaço de tempo. Para remover estes ruídos, em todos os casos em que existiam séries de eventos idênticos em sequência, foi mantido apenas o primeiro evento de cada série.

- *Converting Hidden Complex Events into One* - Dentre os eventos possíveis de serem realizados pelos usuários, alguns obrigatoriamente seguem uma sequência predefinida, o que introduz um ruído que pode levar os algoritmos a identificarem padrões óbvios como relevantes, quando estes são apenas o caminho natural para a ação final a ser realizada. Na base utilizada, este problema foi identificado nos eventos antecedentes à submissão ou tentativa de realização da atividade, pois antes de realizar uma tentativa, o usuário precisa obrigatoriamente visualizar a atividade, e antes de submeter, precisa ter visualizado ou tentado realizar a atividade. Desse modo, foram realizadas as simplificações listadas na Tabela 3.

Tabela 3. Regras utilizadas para conversão de eventos complexos em simples

Subsequência original	Simplificação
assignment_vis>assignment_try	assignment_try
assignment_vis>assignment_sub	assignment_sub
assignment_try>assignment_sub	assignment_sub

A Tabela 4 descreve as oito bases de dados (daqui em diante chamadas de "cenários") resultantes da aplicação das três técnicas de pré-processamento descritas anteriormente de forma individual e combinada, bem como o impacto no tamanho mínimo, médio, mediano, e máximo das sequências em cada cenário. Para geração das bases de cada cenário, os algoritmos de pré-processamento foram implementados em Python como uma API genérica, que pode ser utilizada futuramente por outros pesquisadores, disponível no link `https://github.com/AkiraTorres/moodle-spm-api`.

Tabela 4. Cenários de teste definidos

	Estratégias de pré-processamento aplicadas			Sequências resultantes		
Cenário	Multilevel Sequential Patterns	Coalescing Repeating Point Events into One	Converting Hidden Complex Events into One	Tamanho menor sequência	Tamanho médio das sequências	Tamanho maior sequência
1	X	X	X	1	5,30 (+- 4,89)	54
2	X	X	-	3	7,60 (+- 5,20)	60
3	X	-	X	1	9,04 (+- 11,80)	123
4	X	-	-	3	11,34 (+- 12,02)	127
5	-	X	X	1	5,11 (+- 4,80)	54
6	-	X	-	3	7,41 (+- 5,11)	60
7	-	-	X	1	9,04 (+- 11,80)	123
8	-	-	-	3	11,34 (+- 12,02)	127

Além das sequências em si, foram mantidas nas bases para cada cenário um conjunto de informações contextuais de cada aluno: um identificador único anônimo,

a data/hora em que foi registrado cada evento da sequência, a nota obtida na atividade e a nota máxima possível naquela atividade. As oito bases foram armazenadas como arquivos JSON para serem consumidas pelo algoritmo de SPM.

3.3. Modelagem

Na quarta etapa, foi aplicado o algoritmo de SPM em todos os oito cenários de teste. Considerando que, dados os mesmos parâmetros, qualquer algoritmo de SPM produzirá os mesmos resultados (é uma tarefa determinística) [Zhang and Paquette 2023], foi escolhido para este estudo o algoritmo PrefixSpan por ser de fácil utilização na linguagem Python e ter velocidade de execução superior a outras opções como o algoritmo GSP [Pei et al. 2001].

Assim, o PrefixSpan foi utilizado com suporte mínimo de 0,8 (8%), tamanho mínimo de subsequência a ser encontrada 2, e sem limite máximo de tamanho de subsequência. Também não foi limitado o "gap" entre eventos, ou seja, a distância máxima entre dois eventos numa sequência. Isso permitiu que padrões sequenciais não estritos também fossem identificados, como por exemplo, um grupo de alunos que embora tenham seguido estratégias não idênticas, apresentaram sub-sequências de eventos similares, como um acesso aos recursos de aprendizagem logo no início do prazo seguido por uma participação no fórum no final do prazo da atividade.

3.4. Avaliação

Por fim, os padrões identificados em cada cenário foram avaliados por meio do cálculo de um conjunto de métricas propostas na literatura, além de outras relevantes para este estudo, listadas a seguir:

- **Tamanho do padrão**: número de eventos no padrão identificado. Demonstra se o padrão é mais simples ou se contém uma sequência complexa de ações. [Zhang and Paquette 2023]
- **Suporte**: número total de alunos que exibiram o padrão em sua sequência de aprendizagem [Zhang and Paquette 2023]. Reflete o quanto aquele padrão foi comum no comportamento dos(as) estudantes.
- **I-support**: O total de vezes que o padrão se repete nas sequências de aprendizagem de todos(as) os/as estudantes (um padrão pode ocorrer mais de uma vez em cada sequência) [Lo et al. 2008, Zhang and Paquette 2023]. Ao contrário da métrica de suporte, ao avaliar também a repetição dentro das sequências, é possível identificar padrões que ocorrem de forma recorrente nas trajetórias de aprendizagem dos(as) estudantes.
- **Média das notas obtidas** na atividade pelos(as) estudantes que apresentaram cada um dos padrões identificados. Esta métrica é utilizada para buscar relações entre os padrões de aprendizagem e o desempenho obtido na atividade.
- **Tempo de execução** do algoritmo em cada cenário de teste.

Além disso, foi realizada uma análise qualitativa dos 5 padrões com maior suporte em cada cenário a fim de identificar as diferenças semânticas entre os padrões e sua possível relação com as métricas definidas.

Tabela 5. Métricas obtidas para cada cenário

Cenário	Total de Padrões Identificados	Tamanho Máximo dos Padrões	Tamanho Médio dos Padrões	Média de I-support	Média de Suporte	Tempo de Execução (s)
1	99	4	2,45 (± 0,81)	1.231,12 (± 2.154,84)	15,6% (± 12,3%)	21,4
2	363	6	3,38 (± 0,99)	2.884,76 (± 7.253,16)	16,2% (± 13,4%)	117,0
3	194	8	3,08 (± 1,13)	1.510,58 (± 3.237,57)	14,1% (± 10,0%)	56,3
4	645	9	3,95 (± 1,25)	3.111,46 (± 9.014,66)	14,4% (± 11,0%)	245,9
5	143	5	3,26 (± 0,99)	2.224,58 (± 4.373,78)	16,3% (± 14,0%)	29,9
6	639	7	4,36 (± 1,11)	4.978,21 (± 14.161,66)	15,7% (± 14,3%)	254,8
7	429	12	4,65 (± 1,73)	3.214,92 (±8.449,42)	13,7% (± 10,0%)	157,7
8	1580	13	5,56 (± 1,72)	6.274,82 (± 21.590,31)	13,4% (± 10,4%)	1122,4

4. Resultados

Nesta seção, são apresentados e discutidos os resultados quantitativos (Subseção 4.1) e qualitativos (Subseção 4.2) da aplicação do algoritmo PrefixSpan aos 8 cenários de teste descritos na seção anterior, considerando as métricas de avaliação escolhidas.

4.1. Análise quantitativa

A Tabela 5 resume as métricas obtidas para cada um dos cenários. Pode-se notar que a aplicação conjunta das três estratégias de pré-processamento (cenário 1) reduz sensivelmente a quantidade e tamanho máximo e médio dos padrões identificados pelo algoritmo de SPM, bem como o tempo de execução do algoritmo. A redução da quantidade de padrões ajuda a potencialmente eliminar ruídos na saída do algoritmo, melhorando a qualidade e relevância dos padrões identificados [Zhang and Paquette 2023]. Isto é reforçado pelo aumento do valor médio de suporte em quase todos os cenários em que alguma estratégia de pré-processamento foi aplicada em relação ao cenário 8, onde a base original foi utilizada.

Dentre as três estratégias de pré-processamento utilizadas, a que demonstrou ter maior impacto na métrica suporte foi "*Coalescing Repeating Point Events into One*" (cenários 1, 2, 5 e 6). Isto é sugerido pelo valores de suporte inferiores quando as duas outras estratégias são aplicadas isoladamente (cenários 4 e 7) bem como quando são aplicadas em conjunto (cenário 3). Este resultado indica que no contexto de logs de AVAs, a maior fonte de ruído nos dados para algoritmos de SPM pode ser a ocorrência de eventos idênticos em série, fruto da forma como os dados são capturados pelos sistemas de log dos AVAs.

Já as estratégias *Multilevel Sequential Patterns* e *Converting Hidden Complex Events into One*, tanto quando aplicadas ambas em conjunto (cenário 3) quanto separadamente (cenários 4 e 7), levaram a valores de suporte menores. Em especial, a estratégia *Converting Hidden Complex Events into One*, quando aplicada separadamente (cenário 7), apresentou um suporte quase idêntico ao cenário 8, com a base original. Ainda assim, esses resultados sugerem que a estratégia *Multilevel Sequential Patterns* é ligeiramente mais efetiva em melhorar o nível de suporte que *Converting Hidden Complex Events into One*, já que sua aplicação isolada (cenário 4) levou a um suporte maior que quando aplicada em conjunto com a última (cenário 3).

Por fim, com relação à métrica i-support, observa-se que quanto menor a quanti-

dade de padrões identificados, menor foi o valor obtido para a mesma. Isto sugere que a incidência de repetições de um padrão nas subsequências não foi necessariamente influenciado pelas estratégias de pré-processamento aplicadas.

4.2. Análise qualitativa

De forma a compreender que padrões foram encontrados pelo algoritmo em cada cenário, são listados na Tabela 6 os cinco padrões com maior nível de suporte por cenário. Nos cenários em que a estratégia *Converting Hidden Complex Events into One* não foi aplicada (2, 4, 6, 8), percebe-se que os principais padrões identificados são justamente os eventos complexos que foram simplificados, como *[assignment_try>assignment_sub]* e *[assignment_vis>assignment_try>assignment_sub]*, demonstrando como essa estratégia reduziu o número de padrões irrelevantes.

Já com relação à estratégia *Multilevel Sequential Patterns* (cenários 1-4), é possível perceber um ganho na expressividade dos padrões identificados. Por exemplo, o padrão *[resource_vis_START >assignment_sub_END]* (cenário 3) representa estudantes que acessaram os recursos disponibilizados no início do prazo e entregaram a atividade no final, obtendo uma média superior aos que acessaram os recursos no final do prazo *[resource_vis_END >assignment_sub_END]*. Por fim, a estratégia *Coalescing Repeating Point Events into One*, utilizada nos cenários 1, 2, 5 e 6 foi eficaz em eliminar padrões repetitivos como *[resource_vis >resource_vis]*.

A partir desta análise qualitativa, é possível observar que, apesar do baixo impacto no nível de suporte das estratégias *Multilevel Sequential Patterns* e *Converting Hidden Complex Events into One*, ambas também contribuem para melhorar os resultados dos padrões encontrados pelos algoritmos de SPM.

5. Conclusões e Trabalhos Futuros

Este estudo analisou o impacto de três estratégias de pré-processamento de sequências de eventos em uma base de dados de logs de AVAs em um curso técnico à distância. As estratégias foram testadas com o algoritmo PrefixSpan em oito cenários de forma isolada e combinada e os padrões encontrados foram avaliados utilizando cinco métricas: tamanho do padrão, suporte, i-support, médias das notas obtidas e tempo de execução. Os resultados desta avaliação sugerem que a estratégia "*Coalescing Repeating Point Events into One*" teve o maior impacto no nível médio de suporte, mas as outras duas, *Multilevel Sequential Patterns* e *Converting Hidden Complex Events into One* também contribuem para melhorar a qualidade dos padrões detectados. A principal limitação do estudo foi a utilização de dados de um único curso na avaliação das estratégias de pré-processamento. Como trabalhos futuros, sugere-se a avaliação com outras bases de dados bem como a variação de mais parâmetros suportados pelos algoritmos de SPM como padrões fechados, padrões geradores, e diferentes valores de *gap*.

6. Agradecimentos

José Thiago Torres da Silva possui bolsa de Iniciação Científica financiada pelo programa UPE-PFA (Edital ICTI 2023, processo P00226/S00226).

Tabela 6. Lista dos cinco padrões com maior nível de suporte, por cenário

Cenário	Tamanho do padrão	Padrão	I-suporte	Suporte	Média de Notas
1	2	resource_vis_END>assignment_sub_END	5657	62,44%	1,4
1	2	resource_vis_START>assignment_sub_END	3734	35%	1,42
1	2	assignment_try_END>assignment_sub_END	2176	32,85%	1,38
1	2	assignment_vis_END>assignment_sub_END	3054	29,64%	1,4
1	2	forum_vis_END>assignment_sub_END	4042	27%	1,41
2	2	assignment_vis_END>assignment_sub_END	37665	90,02%	1,38
2	2	assignment_vis_END>assignment_try_END	26634	81,53%	1,37
2	2	assignment_try_END>assignment_sub_END	23240	81,53%	1,37
2	3	assignment_vis_END>assignment_try_END>assignment_sub_END	13317	81,53%	1,37
2	2	resource_vis_END>assignment_sub_END	21351	62,44%	1,4
3	2	resource_vis_END>assignment_sub_END	10310	62,44%	1,4
3	2	assignment_vis_END>assignment_sub_END	8228	45,41%	1,38
3	2	resource_vis_START>assignment_sub_END	6398	35%	1,42
3	2	assignment_try_END>assignment_sub_END	2704	32,85%	1,38
3	2	resource_vis_END>resource_vis_END	6130	30,31%	1,37
4	2	assignment_vis_END>assignment_sub_END	59475	88,44%	1,38
4	2	assignment_vis_END>assignment_try_END	37760	81,53%	1,37
4	2	assignment_try_END>assignment_sub_END	32339	81,53%	1,37
4	3	assignment_vis_END>assignment_try_END>assignment_sub_END	18880	81,53%	1,37
4	2	resource_vis_END>assignment_sub_END	33156	62,44%	1,4
5	2	resource_vis>assignment_sub	15585	87,72%	1,41
5	2	assignment_try>assignment_sub	4925	46,71%	1,38
5	2	assignment_vis>assignment_sub	8714	43,26%	1,39
5	2	assignment_vis>resource_vis	8052	39,8%	1,39
5	3	assignment_vis>resource_vis>assignment_sub	4026	39,8%	1,39
6	2	assignment_vis>assignment_try	70735	100%	1,38
6	2	assignment_vis>assignment_sub	83018	100%	1,38
6	2	assignment_try>assignment_sub	58468	100%	1,38
6	3	assignment_vis>assignment_try>assignment_sub	36878	100%	1,38
6	2	resource_vis>assignment_sub	68164	87,72%	1,41
7	2	resource_vis>assignment_sub	41663	87,72%	1,41
7	2	assignment_vis>assignment_sub	26501	59,37%	1,37
7	2	resource_vis>resource_vis	38924	52,85%	1,4
7	3	resource_vis>resource_vis>assignment_sub	19462	52,85%	1,4
7	2	assignment_try>assignment_sub	8816	46,71%	1,38
8	2	assignment_vis>assignment_try	132929	100%	1,38
8	2	assignment_vis>assignment_sub	171556	100%	1,38
8	2	assignment_try>assignment_sub	107417	100%	1,38
8	3	assignment_vis>assignment_try>assignment_sub	68416	100%	1,38
8	2	resource_vis>assignment_sub	150501	87,72%	1,41

Referências

[Agrawal and Srikant 1995] Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, pages 3–14.

[Azevedo and Santos 2008] Azevedo, A. and Santos, M. F. (2008). KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW. pages 182–185.

[Baker 2014] Baker, R. S. (2014). Educational data mining: An advance for intelligent systems in education. *IEEE Intelligent systems*, 29(3):78–82.

[Bogarín et al. 2018] Bogarín, A., Cerezo, R., and Romero, C. (2018). A survey on educational process mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(1):e1230.

[Calders and Pechenizkiy 2012] Calders, T. and Pechenizkiy, M. (2012). Introduction to the special section on educational data mining. *Acm Sigkdd Explorations Newsletter*, 13(2):3–6.

[Chen et al. 2020] Chen, G., Rolim, V., Mello, R. F., and Gašević, D. (2020). Let's shine together! a comparative study between learning analytics and educational data mining. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, LAK '20, pages 544–553, New York, NY, USA. Association for Computing Machinery.

[Du et al. 2017] Du, F., Shneiderman, B., Plaisant, C., Malik, S., and Perer, A. (2017). Coping with Volume and Variety in Temporal Event Sequences: Strategies for Sharpening Analytic Focus. *IEEE Transactions on Visualization and Computer Graphics*, 23(6):1636–1649. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[Fournier Viger et al. 2017] Fournier Viger, P., lin, j., Rage, U., Koh, Y. S., and Thomas, R. (2017). A Survey of Sequential Pattern Mining. *Data Science and Pattern Recognition*, 1:54–77.

[Guo et al. 2022] Guo, Y., Guo, S., Jin, Z., Kaul, S., Gotz, D., and Cao, N. (2022). Survey on Visual Analysis of Event Sequence Data. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):5091–5112. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[Han and Kamber 2012] Han, J. and Kamber, M. (2012). *Data mining: concepts and techniques*. Elsevier, Burlington, MA, 3rd ed edition.

[Lo et al. 2008] Lo, D., Khoo, S.-C., and Liu, C. (2008). Efficient mining of recurrent rules from a sequence database. In *Database Systems for Advanced Applications: 13th International Conference, DASFAA 2008, New Delhi, India, March 19-21, 2008. Proceedings 13*, pages 67–83. Springer.

[Lockyer et al. 2013] Lockyer, L., Heathcote, E., and Dawson, S. (2013). Informing Pedagogical Action: Aligning Learning Analytics With Learning Design. *American Behavioral Scientist*, 57(10):1439–1459.

[Maranhão et al. 2023] Maranhão, D., Borges, P., and Neto, C. (2023). Descoberta de padrões sequenciais de aprendizagem em um ambiente voltado ao ensino de algorit-

mos. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 1385–1396, Porto Alegre, RS, Brasil. SBC.

[Munk et al. 2017] Munk, M., Drlík, M., Benko, L., and Reichel, J. (2017). Quantitative and qualitative evaluation of sequence patterns found by application of different educational data preprocessing techniques. *IEEE Access*, 5:8989–9004.

[Pei et al. 2001] Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., and Hsu, M.-C. (2001). Prefixspan,: mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings 17th International Conference on Data Engineering*, pages 215–224.

[Peña-Ayala 2023] Peña-Ayala, A. (2023). *Educational Data Science: Essentials, Approaches, and Tendencies*. Springer.

[Poon et al. 2017] Poon, L. K., Kong, S.-C., Wong, M. Y., and Yau, T. S. (2017). Mining sequential patterns of students' access on learning management system. In *Data Mining and Big Data: Second International Conference, DMBD 2017, Fukuoka, Japan, July 27–August 1, 2017, Proceedings 2*, pages 191–198. Springer.

[Schröer et al. 2021] Schröer, C., Kruse, F., and Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, 181:526–534.

[Song et al. 2022] Song, W., Ye, W., and Fournier-Viger, P. (2022). Mining sequential patterns with flexible constraints from mooc data. *Applied Intelligence*, 52(14):16458–16474.

[Verbert et al. 2020] Verbert, K., Ochoa, X., De Croon, R., Dourado, R. A., and De Laet, T. (2020). Learning analytics dashboards: the past, the present and the future. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, LAK '20, pages 35–40, New York, NY, USA. Association for Computing Machinery.

[Wise 2019] Wise, A. F. (2019). Learning Analytics: Using Data-Informed Decision-Making to Improve Teaching and Learning. In Adesope, O. O. and Rud, A., editors, *Contemporary Technologies in Education*, pages 119–143. Springer International Publishing, Cham.

[Wise and Jung 2019] Wise, A. F. and Jung, Y. (2019). Teaching with Analytics: Towards a Situated Model of Instructional Decision-Making. *Journal of Learning Analytics*, 6(2):53–69–53–69. Number: 2.

[Zaki 2000] Zaki, M. J. (2000). Sequence mining in categorical domains: incorporating constraints. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*, CIKM '00, page 422–429, New York, NY, USA. Association for Computing Machinery.

[Zhang and Paquette 2023] Zhang, Y. and Paquette, L. (2023). Sequential Pattern Mining in Educational Data: The Application Context, Potential, Strengths, and Limitations. In Peña-Ayala, A., editor, *Educational Data Science: Essentials, Approaches, and Tendencies: Proactive Education based on Empirical Big Data Evidence*, pages 219–254. Springer Nature, Singapore.

[Zhou et al. 2010] Zhou, M., Xu, Y., Nesbit, J., and Winne, P. (2010). *Sequential Pattern Analysis of Learning Logs*, pages 107–121.