Um algoritmo genético com função de aptidão flexível para seleção de atributos em dados educacionais*

Danielle F. de Albuquerque, Diego N. Brandão, Rafaelli Coutinho

[1]Programa de Pós-graduação em Ciência da Computação
Centro Federal de Educação Tecnológica Celso Suckow da Fonseca – CEFET/RJ
Rio de Janeiro – RJ – Brasil
danielle.albuquerque@aluno.cefet-rj.br
{diego.brandao, rafaelli.coutinho}@cefet-rj.br

Abstract. *Due to the growing volume and increasing availability of educational data, data mining techniques have been frequently applied to help understand phenomena related to education. However, much of this data can be sparse, redundant, irrelevant, and noisy, which can degrade the quality and the computational performance of predictive models. One way to minimize these problems is to select attributes in the modeling process using Feature Selection (FS) techniques. This article proposes a FS approach with a genetic algorithm adapted to the educational context. The results indicate that the proposal improves classification performance and allows education specialists to have greater flexibility in selecting attributes according to their needs and realities.*

Resumo. *Devido ao volume crescente e à disponibilidade cada vez maior de dados educacionais, as técnicas de mineração de dados têm sido frequentemente aplicadas para auxiliar na compreensão do desempenho escolar do aluno. No entanto, muitos desses dados podem ser esparsos, redundantes, irrelevantes e com ruído, o que pode prejudicar a qualidade e o desempenho computacional dos modelos preditivos. Uma maneira de minimizar esses problemas consiste em selecionar atributos no processo de modelagem por meio de técnicas de Seleção de Atributos (SA). Este artigo propõe uma abordagem de SA com algoritmo genético adaptada ao contexto educacional. Os resultados indicam que a proposta melhora o desempenho da classificação e permite que especialistas em educação tenham uma maior flexibilidade no processo de seleção dos atributos de acordo com suas necessidades e realidades.*

1. Introdução

O advento da Internet e a digitalização dos processos levaram o cenário educacional para uma nova realidade nos últimos anos. Cursos no formato a distância, *softwares* educacionais, bancos de dados públicos e sistemas de gestão escolar computadorizados são alguns exemplos das tecnologias que criaram grandes repositórios de dados capazes de gerar informações que descrevem a jornada de estudantes. Neste contexto, emerge a área de Mineração de Dados Educacionais (MDE) para auxiliar na compreensão e análise de tais dados. Dentre os objetivos da MDE, destaca-se o fornecimento de ferramentas que possibilitem o mapeamento do perfil de estudantes, do risco de evasão acadêmica e dos

*Os autores agradecem à CAPES (código 001) pelo financiamento do projeto.

principais fatores que impactam no desempenho escolar, permitindo que o investimento financeiro e pedagógico seja feito de maneira eficiente [Romero and Ventura, 2010].

Assim como em outros domínios, os dados educacionais também podem ser esparsos, redundantes, irrelevantes e com ruído, o que pode prejudicar a qualidade e o desempenho computacional dos modelos exploratórios e preditivos [Farissi et al., 2020]. Uma maneira de minimizar esses problemas consiste em identificar informações menos importantes e omiti-las durante o processo de modelagem por meio de técnicas de Seleção de Atributos (SA).

Diversas técnicas de SA têm sido usadas na área da educação. Muitos trabalhos utilizam métodos de filtro com critérios estatísticos de ranqueamento devido a facilidade do seu uso [Febro, 2019]. Outros utilizam abordagem *wrapper* com o uso de algoritmo genético (AG) para selecionar o melhor conjunto de atributos que definem o cenário escolar e acadêmico [Farissi et al., 2020; Santos et al., 2020], com maior custo computacional. Outros ainda aplicam o próprio classificador em uma abordagem denominada de embutida [Gitinabard et al., 2018]. Além disso, comparações dessas abordagem também já foram realizadas a fim de compreender qual delas é mais adequada ao cenário a ser estudado [Ahmed et al., 2020].

Os trabalhos encontrados na literatura apresentam apenas técnicas clássicas de SA, sem propor uma abordagem específica para dados educacionais. Nesse sentido, este artigo propõe uma nova abordagem de SA com algoritmo genético adaptado ao contexto educacional. Ela possibilita que especialistas em educação tenham maior flexibilidade para inserirem informações empíricas no processo de SA de acordo com as diversas necessidades e realidades que o ambiente educacional apresenta, permitindo que características intrínsecas de cada ambiente escolar sejam melhor exploradas.

2. Fundamentação Teórica

A SA é uma etapa do pré-processamento dos dados, onde atributos são removidos no processo de modelagem, de forma a selecionar um subconjunto que representará melhor a base de dados. Diversos benefícios podem ser alcançados a partir desse processo, como: tornar os algoritmos mais rápidos e eficientes, melhorar a precisão da classificação, melhorar a qualidade dos dados, evitar o *overfitting* e facilitar a visualização dos resultados [Chandrashekar and Sahin, 2014].

Apesar de todos os benefícios, a SA é um problema complexo, fazendo parte da classe dos problemas NP [Davies and Russell, 1994; Chen et al., 1997]. Por isso, diversos métodos de SA foram propostos, como AG e métodos de filtro [Tan et al., 2008; Febro, 2019] ambos utilizados na metodologia deste artigo.

O AG é uma meta-heurística baseada no princípio da evolução. As soluções candidatas, chamadas de indivíduos, são avaliadas por uma função objetivo, conhecida como função de aptidão, de forma que a mais apta "sobreviva" para a próxima geração. A cada geração são realizadas recombinações dos indivíduos mais aptos por meio das operações de mutação e cruzamento, formando uma nova população de indivíduos para a próxima geração. Esse procedimento é feito de forma iterativa até que um critério de parada seja alcançado, por exemplo atingir um determinado número de gerações ou atingir determinado valor na função objetivo [Tan et al., 2008; Santos et al., 2020].

Já os métodos de filtro costumam ser mais simples e rápidos, guiados por uma medida estatística de avaliação que dispõe os atributos em ordem decrescente de importância. Dentre estes métodos, o Qui-Quadrado (QQ) é uma técnica bastante utilizada em dados educacionais [Febro, 2019]. O cálculo do QQ, representado pela equação $QQ = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}}$, utiliza as frequências observadas e estimadas entre o atributo e a classe, onde n é a quantidade de categorias possíveis para aquele atributo, k é a quantidade de classes, A_{ij} é a frequência observada da categoria i na classe j e E_{ij} é a frequência estimada, calculada pela distribuição dos dados no atributo e na classe.

3. Metodologia

A metodologia adotada, ilustrada na Figura 1, consiste em três etapas principais: 1) pré-processamento dos dados; 2) seleção de atributos usando um AG com função de aptidão flexível para dados educacionais e incorporado a uma abordagem híbrida com QQ + AG; e 3) classificação.

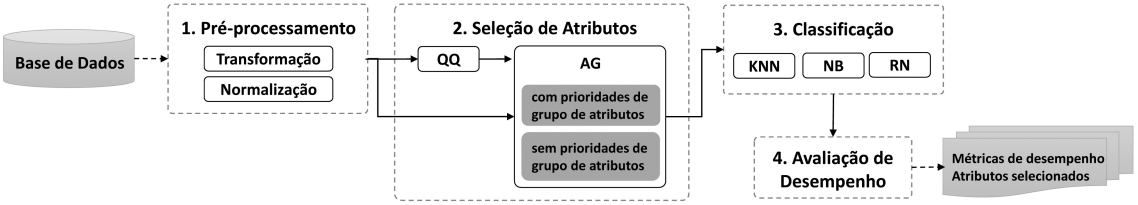

Figura 1. Visão geral da metodologia adotada.

O conjunto de dados usado possui 480 estudantes e 16 atributos de uma plataforma escolar de *e-learning*[1] [Zaffar et al., 2018; Jalota and Agrawal, 2021; Farissi et al., 2020]. O atributo-alvo da classificação é a `Classe`, que representa o desempenho do aluno, podendo este ser baixo, médio ou alto. Os atributos são divididos em três grupos principais: (1) Características comportamentais: mão levantada na classe, quantidade de visitas no conteúdo do site, número de visualizações de notícias, participação nos grupos de discussão, resposta de pesquisas de satisfação pelos pais, grau de satisfação e quantidade de faltas. (2) Características demográficas: gênero, local de nascimento e nacionalidade; e (3) Características de formação acadêmica: estágio educacional, nível de série, turma, disciplina e semestre.

A etapa de pré-processamento inclui a transformação de dados categóricos e a normalização dos dados numéricos. Os atributos categóricos são convertidos em binários resultando no aumento de 16 para 70 atributos. Essa etapa foi necessária para que os valores referentes às categorias não fossem considerados como uma escala numérica [Hancock, 2020]. Para evitar algoritmos enviesados, a normalização de dados numéricos foi realizada por meio do método MinMax[2], com os valores sendo alterados para o intervalo de 0 a 1.

A etapa seguinte consiste na seleção otimizada de atributos para o contexto educacional. O objetivo é investigar se grupos de atributos são mais relevantes que outros na análise de desempenho de estudantes e a partir disso, propor um método de seleção

[1]Disponível no repositório Kaggle: https://www.kaggle.com/aljarah/xAPI-Edu-Data

[2]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

de atributos com AG que priorize esses grupos. Para isso, é necessário que haja uma categorização dos atributos em grupos de acordo com a natureza deles, conforme a base de dados utilizada que divide os atributos em comportamentais, demográficos e acadêmicos.

O AG seleciona os atributos criando conjuntos que serão avaliados pela função objetivo FO proposta na Equação 1. Cada conjunto é um indivíduo representado por um vetor binário b, onde cada elemento b_i indica a presença (1) ou não (0) de um atributo i e $|b| = N$, número total de atributos do conjunto de dados. A cada geração acontecem as operações de cruzamento e mutação, que fazem a variação evolutiva de cada indivíduo. As prioridades dos grupos de atributos são atribuídas por meio de ponderamentos (pesos) na FO a ser maximizada. A função é composta por três componentes: 1) a acurácia da classificação com os atributos selecionados (acc); 2) uma penalização referente ao número de atributos selecionados (num_select); e 3) uma parcela com os ponderamentos das prioridades dos grupos de atributos, *i.e.*, o somatório da multiplicação do peso do atributo i (p_i) por b_i.

$$FO = acc - \frac{num\_select}{N} + \frac{1}{\gamma} \times \sum_{i=1}^{N} p_i \times b_i \tag{1}$$

onde $\gamma = \sum_{i=1}^{N} p_i$ é um fator de normalização.

Essa abordagem fornece aos especialistas em educação uma ferramenta flexível que lhes permite atribuir as prioridades dos grupos de atributos de acordo com a realidade de cada ambiente escolar. Por exemplo, no ensino presencial atributos de localização poderiam ser intuitivamente mais relevantes, o que pode ser avaliado com a abordagem flexível proposta. Nesse caso, especialistas em educação poderiam dar um peso maior para o grupo dos atributos demográficos, em relação aos do tipo comportamental. O AG proposto também foi incorporado a uma abordagem híbrida, que aplica o método clássico do QQ antes do AG, de maneira a potencializar o ganho de desempenho da classificação [Singh and Selvakumar, 2015]. Além disso, para fins de comparação, um AG com a FO sem a parcela de prioridades foi avaliado. Para a última etapa, os classificadores utilizados foram o *K-Nearest Neighbors* (KNN), o Naive Bayes (NB) e a Rede Neural (RN), pois tais técnicas apresentaram resultados mais interessantes na literatura [Zaffar et al., 2018; Farissi et al., 2020].

4. Resultados

Os experimentos foram realizados na plataforma Google Colab com $13,3$ GB de memória RAM, Intel(R) Xeon(R) CPU @ 2.30GHz, e implementados em Python[3] usando as bibliotecas *pandas*, *scikit-learn*, *prince* e *deap*[4]. No AG foram utilizados os seguintes parâmetros: população $= 30$, geração $= 30$, $crossover = 0,09$ e mutação $= 0,01$. A RN escolhida foi a *Multilayer Perceptron* (MLP) com um máximo de 300 interações. O NB foi o gaussiano com $smoothing = 0,01$. No KNN foi utilizada a distância euclidiana com $k = 5$ e $p = 2$. Os demais parâmetros de AG e classificadores foram os valores padrões das funções nas bibliotecas. A base de dados foi dividida em 70% para

[3]Disponível em: https://github.com/SCICOM-CEFET-RJ/Educational-DataMining

[4]https://pandas.pydata.org, https://scikit-learn.org/stable, https://pypi.org/project/prince e https://deap.readthedocs.io/en/master

treinamento e 30% para teste. Para a avaliação dos modelos de classificação, a métrica f_1 foi usada, uma vez que consiste da média harmônica de outras métricas, precisão e *recall*.

Inicialmente, o AG foi avaliado com uma função de aptidão constituída apenas da acurácia do classificador e da penalização do número de atributos selecionados. Observou-se que alguns atributos estavam mais presentes nos resultados independente dos parâmetros e classificadores utilizados. Assim, decidiu-se avaliar também a importância dos atributos ranqueando-os previamente por meio do método QQ[5]. O resultado mostrou que os atributos do grupo comportamental estão presentes nas primeiras posições e que os atributos referentes às características acadêmicas aparecem menos nessa lista. Dessa forma, os atributos que refletem as atitudes dos alunos dentro da plataforma parecem ser mais relevantes para definir o desempenho deles neste cenário de educação a distância. Tal observação permite criar a hipótese de que priorizar determinados grupos de atributos, mais relacionados ao contexto em que os alunos estão inseridos, pode ajudar na seleção otimizada de características que contribuem para análise de desempenho escolar.

Com base nisso, o AG proposto foi avaliado priorizando os atributos de cada grupo da base de dados (C: Comportamental, D: Demográfica e A: Acadêmica) e na abordagem híbrida, QQ + AG. De forma alternada, um grupo de atributos recebe uma prioridade maior de $0,70$, enquanto que os demais, uma mesma prioridade de $0,15$. A Tabela 1 apresenta os resultados para os experimentos sem SA, AG com a função de aptidão sem prioridade de grupos, AG com a função de aptidão proposta, e a abordagem híbrida QQ + AG, usando os 40 primeiros atributos do ranqueamento do QQ. Para cada classificador, são apresentados a média e o desvio padrão da métrica f_1 para 10 execuções, e a quantidade média de atributos selecionados.

SA	**KNN**		**NB**		**RN**	
	f_1	**# atr.**	f_1	**# atr.**	f_1	**# atr.**
Sem SA	67,1 $\pm$ 4,3	70	63,8 $\pm$ 4,6	70	72,6 $\pm$ 3,2	70
AG	70,8 $\pm$ 3,5	21	64,0 $\pm$ 4,3	22	72,4 $\pm$ 3,3	21
AG (C)	77,2 $\pm$ 3,1	32	69,0 $\pm$ 2,7	31	75,7 $\pm$ 2,4	32
AG (D)	73,2 $\pm$ 5,0	37	65,5 $\pm$ 3,0	34	74,0 $\pm$ 3,1	36
AG (A)	73,9 $\pm$ 3,5	35	66,0 $\pm$ 4,5	34	73,0 $\pm$ 4,2	38
QQ + AG	70,9 $\pm$ 3,8	10	71,0 $\pm$ 4,0	9	71,1 $\pm$ 5,0	9
QQ + AG (C)	76,2 $\pm$ 3,3	18	72,1 $\pm$ 2,4	18	76,0 $\pm$ 1,8	17
QQ + AG (D)	72,7 $\pm$ 4,0	21	67,0 $\pm$ 3,9	19	73,0 $\pm$ 3,8	22
QQ + AG (A)	70,0 $\pm$ 6,8	19	70,8 $\pm$ 4,8	20	72,2 $\pm$ 3,7	20

Tabela 1. Resultados dos experimentos

Todos os classificadores apresentaram ganhos de desempenho com as abordagens de SA em relação ao modelo sem SA. Na avaliação do AG com grupos prioritários, os melhores resultados de média foram obtidos quando os atributos comportamentais foram priorizados. Em relação ao uso da abordagem híbrida, os resultados com o grupo comportamental como prioridade também se mostraram mais assertivos na média do que o alcançado pelo modelo sem uso das prioridades. Isso confirma a hipótese feita anteriormente e indica que uma função de aptidão flexível pode auxiliar a compreensão da diversidade de realidades do cenário educacional. Sobre a quantidade de atributos selecionada, ela foi em média menor quando o AG sem prioridades foi usado e se manteve na faixa de 35 com apenas o AG proposto e 20 quando incorporado à abordagem híbrida.

[5]Selecionou-se os 40 primeiros atributos com maior valor no ranqueamento QQ, pois este conjunto de atributos apresentou maior f_1 do que os outros valores avaliados: 10, 20 ou 30.

O melhor f_1 obtido nos experimentos, de $77,2$, foi na abordagem AG (C) com o KNN, o que resultou na seleção de 32 atributos. No entanto, ao analisar o desvio padrão dos experimentos, os resultados são estatisticamente semelhantes.

5. Considerações finais

A seleção de atributos é uma técnica importante e tem sido cada vez mais usada para identificar os principais fatores que impactam no desempenho do aluno no contexto da mineração de dados educacionais. Este artigo apresentou uma abordagem adaptável da função objetivo do AG para seleção de atributos em dados educacionais no contexto do *e-learning*. É possível afirmar que o AG proposto aumentou em média a assertividade da seleção de atributos e consequentemente melhorou o desempenho de classificação. No cenário de educação *e-learning*, foi observado que os atributos do tipo comportamental são relevantes para melhorar a qualidade da classificação. No entanto, no ensino presencial, outros atributos, como os demográficos, poderiam ser intuitivamente mais relevantes, o que pode ser avaliado com a abordagem flexível proposta. Isso se aplica ao contexto brasileiro onde a diversidade demográfica e social são tão presentes na realidade escolar. Assim, pretende-se, como trabalho futuro, explorar as técnicas em outras bases educacionais abertas para avaliar diversos ambientes educacionais, como a base disponibilizada pelo governo brasileiro sobre o censo de educação superior.

Referências

M.R. Ahmed et al. A comprehensive analysis on undergraduate student academic performance using feature selection techniques on classification algorithms. In *ICCCNT 2020*, pages 1–6, 2020.

G. Chandrashekar and F. Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

B. Chen, J. Hong, and Y. Wang. The minimum feature subset selection problem. *Journal of Computer Science and Technology*, 12(2):145–153, 1997.

S. Davies and S. J. Russell. NP-completeness of searches for smallest possible feature sets. In *AAAI Symposium on Intelligent Relevance*, pages 37–39. AAAI Press, 1994.

A. Farissi et al. Genetic algorithm based feature selection for predicting student's academic performance. *Emerging Trends in Intelligent Computing and Informatics*, pages 110–117, 2020.

J. D Febro. Utilizing feature selection in identifying predicting factors of student retention. *International Journal of Advanced Computer Science and Applications*, 10(9), 2019.

N. Gitinabard et al. Your actions or your associates? predicting certification and dropout in moocs with behavioral and social features. In *EDM*, 2018.

Khoshgoftaar T.M. Hancock, J.T. Survey on categorical data for neural networks. *Journal of Big Data*, 28(7), 2020.

C. Jalota and R. Agrawal. Feature selection algorithms and student academic performance: A study. *Advances in Intelligent Systems and Computing*, 1165:317–328, 2021.

C. Romero and S. Ventura. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 2010.

G. A.S. Santos et al. EvolveDTree: Analyzing Student Dropout in Universities. *In ICSSIP*, pages 173–178, 2020.

S. Singh and S. Selvakumar. A hybrid feature subset selection by combining filters and genetic algorithm. In *ICCCA*, pages 283–289, 2015.

F. Tan, X. Fu, Y. Zhang, and A. G. Bourgeois. A genetic algorithm-based method for feature subset selection. *Soft Computing*, 12(2):111–120, 2008. ISSN 14327643.

M. Zaffar, M. A. Hashmani, and K. S. Savita. Performance analysis of feature selection algorithm for educational data mining. In *ICBDA 2017*, pages 7–12, 2018.