Árvores de Decisão Aplicadas na Previsão de Desempenho de Alunos: Estado da Arte

Marcos Alves Vieira[1], **Ernesto Fonseca Veiga**[1]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)
Goiânia – GO – Brasil

{marcosalves, ernestofonseca}@inf.ufg.br

Abstract. *Educational Data Mining (EDM) is a field that uses machine learning, data mining, and statistics to process educational data, aiming to reveal useful information for analysis and decision making. This work aims to present a survey of the state of the art through the analysis of relevant and recent articles in the field, focusing on methods that use decision trees for predicting students performance. Additionally, it presents the concepts and basic foundations of EDM.*

Resumo. *A mineração de dados em ambientes educacionais (EDM) é uma área que faz uso de aprendizagem de máquina, mineração de dados e estatística para processar dados educacionais, com objetivo de revelar informações úteis para análise e tomada de decisão. Este trabalho tem a finalidade de apresentar um levantamento do estado da arte, por meio da análise de artigos relevantes e recentes na área, com foco nos que utilizam métodos de árvores de decisão para a previsão de desempenho de alunos. Além disso, também são apresentados os conceitos e fundamentos de EDM.*

1. Introdução

A mineração de dados em ambientes educacionais (*Educational Data Mining* - EDM), é uma área emergente que se vale de técnicas de mineração de dados para revelar informações importantes, presentes em bases de dados de instituições de ensino que, de outra forma, seriam dificilmente perceptíveis. Esta área ganha ainda mais importância nos dias de hoje com a crescente utilização de plataformas de aprendizagem informatizadas, os chamados sistemas de gestão de aprendizagem (*Learning Management Systems* - LMS), inclusive em se tratando de grandes volumes de dados.

A utilização da EDM permite descobrir novos conhecimentos com base em dados gerados pelos alunos, de modo a ajudar a validar e avaliar os sistemas de ensino, visando melhorar alguns aspectos da qualidade da educação, além de estabelecer bases para um processo de aprendizagem mais eficaz. Há ainda outros exemplos de utilização de EDM, tais como, guiar o aprendizado dos estudantes e obter métodos que possam avaliar uma nova metodologia de ensino [Romero and Ventura 2010].

O restante desse trabalho está estruturado como se segue. A Seção 2 aborda os fundamentos básicos de EDM. Na Seção 3, é realizado um levantamento do estado da arte em EDM, analisando trabalhos recentes de maior relevância nesta área. E, finalmente, na Seção 4, são apresentadas as conclusões deste trabalho.

2. Mineração de Dados Educacionais

A quantidade de dados armazenados em bancos de dados educacionais aumenta rapidamente. À medida em que isto acontece, diferentes técnicas de mineração de dados têm sido desenvolvidas e utilizadas com a finalidade de obter informações a partir de tais dados e também para encontrar relações ocultas entre as variáveis utilizadas [Han et al. 2006].

A mineração de dados (*data mining*) é uma ampla área que integra diferentes técnicas, incluindo aprendizado de máquina, estatística, reconhecimento de padrões, inteligência artificial e sistemas de bancos de dados para a análise de grandes volumes de informações [Wu et al. 2008]. Assim sendo, mineração de dados é um processo para extração de padrões previamente desconhecidos, válidos e potencialmente úteis que estão escondidos em grandes conjuntos de dados, conduzindo a um nível incremental de informação e conhecimento [Connolly and Begg 2005, Aviad and Roy 2011].

Por sua vez, a mineração de dados em ambientes educacionais (EDM) é um campo que explora estatística, aprendizado de máquina e algoritmos de mineração de dados aplicados a diferentes tipos de dados de ensino. Seu principal objetivo é analisar estes dados a fim de resolver questões de investigação educacional. A EDM está preocupada com o desenvolvimento de métodos para explorar os dados presentes em ambientes de ensino e, através destes métodos, compreender melhor os estudantes e as condições em que eles aprendem [Baker et al. 2010].

A grande quantidade de dados gerada pode ser atribuída ao crescimento de dois tipos de ferramentas voltadas para o ensino. Por um lado, o aumento tanto de *softwares* educacionais, como dos bancos de dados de informação sobre estudantes, criaram grandes repositórios de dados que refletem em como os alunos aprendem [Koedinger et al. 2008]. Por outro lado, o uso da Internet na educação criou um novo contexto de educação, conhecido como *e-learning*, em que grandes quantidades de informação sobre a interação ensino-aprendizagem são constantemente geradas e disponibilizadas de forma ubíqua [Castro et al. 2007].

O processo de EDM converte os dados brutos, provenientes dos sistemas de ensino, em informações úteis que podem ter um grande impacto na pesquisa e na prática educativa. Este processo não difere muito de outras áreas de aplicação de técnicas de mineração, como negócios, genética, medicina, entre outras, porque segue os mesmos passos do processo geral de mineração de dados [Romero et al. 2004]: pré-processamento, mineração de dados, e pós-processamento.

Do ponto de vista prático, a EDM permite, por exemplo, a descoberta de novos conhecimentos com base em dados de uso dos alunos, a fim de ajudar a validar e avaliar os sistemas educacionais e melhorar potencialmente alguns aspectos da qualidade da educação, além de estabelecer as bases para um processo de aprendizagem mais eficaz [Romero et al. 2004]. Algumas ideias semelhantes já foram utilizadas com sucesso em sistemas *e-commerce*, a primeira e mais popular aplicação de mineração de dados [Raghavan 2005], para determinar os interesses dos clientes e aumentar as vendas.

A EDM envolve diferentes grupos de usuários ou participantes. Diferentes grupos, olham para informações educacionais a partir de ângulos diferentes, de acordo com sua responsabilidade, visão e objetivos para a utilização da mineração dos dados educacionais [Hanna 2004]. Por exemplo, o conhecimento descoberto por algoritmos de EDM

pode ser usado não só para ajudar os professores na gestão de suas aulas, a compreender os processos de aprendizagem de seus alunos e refletir sobre os seus próprios métodos de ensino, mas também para apoiar as reflexões sobre a situação dos alunos e dar um *feedback* aos mesmos [Merceron and Yacef 2005].

Atualmente, há uma grande variedade de sistemas e ambientes de ensino, tais como: a sala de aula tradicional, *e-learning*, LMS, Sistemas Hipermídia Adaptativos, testes/questionários, textos/conteúdos. Outras ferramentas também têm se destacado e contribuído nesse contexto, como por exemplo: objetos de aprendizagem, repositórios, mapas conceituais, redes sociais, fóruns, ambientes de jogos educativos, ambientes virtuais, ambientes de computação ubíqua, etc. Cada um desses sistemas e ambientes educacionais mencionados anteriormente fornece diferentes tipo de dados, permitindo assim, que diversos problemas e tarefas sejam resolvidos através da utilização de técnicas de mineração [Romero and Ventura 2010].

3. Aplicação de Árvores de Decisão em Mineração de Dados Educacionais

Romero e Ventura [Romero and Ventura 2010] mostraram que árvores de decisão é uma abordagem útil em três áreas de EDM. São elas:

1. **Fornecimento de *feedback* para apoiar professores**: tem como objetivo fornecer *feedback* para apoiar os coordenadores de curso, professores e administradores na tomada de decisão, para, por exemplo, melhorar a aprendizagem dos alunos e organizar recursos institucionais de forma mais eficiente. Além disso, permite que os coordenadores tomem medidas proativas apropriadas, apoiadas em dados, e/ou medidas corretivas. Muitos estudos se aplicam a comparar vários modelos de mineração de dados que fornecem *feedback* que podem ser úteis nos objetivos citados, como: regras de associação, *clustering*, classificação e análise de padrões sequenciais.
2. **Recomendações para estudantes**: o objetivo das aplicações nesta área é ser capaz de fazer recomendações diretamente aos alunos no que diz respeito às suas atividades pessoais, *links* para visitas, a próxima tarefa ou exercício a ser feito, entre outras. Outros objetivos são a capacidade de se adaptar a conteúdos de aprendizagem, interfaces e ao progresso de cada aluno em particular. Diversas técnicas de mineração de dados têm sido utilizados para esta tarefa, onde as mais comuns são: de mineração de regras de associação, *clustering*, e mineração de padrões sequenciais. Estas técnicas visam identificar as relações entre as ocorrências de eventos para descobrir se existe alguma ordem específica.
3. **Previsão de desempenho de alunos**: tem como objetivo estimar o valor desconhecido de uma variável que descreve o estudante, de forma particular, voltando-se para o seu desempenho. A previsão de desempenho de alunos é uma das mais antigas e mais populares aplicações de EDM. Diferentes tipos de modelos de redes neurais têm sido utilizados para previsão de notas de estudantes; previsão do número de erros que um aluno cometerá em uma avaliação; e previsão do resultado final de alunos em disciplinas (aprovado ou reprovado).

A seguir é apresentado um levantamento de trabalhos recentes e relevantes em mineração de dados educacionais com emprego de árvores de decisão para previsão de desempenho de alunos.

Tabela 1. Descrição dos atributos e seus possíveis valores

Attribute	Description	Possible Values
Branch	Student's branch in B.Tech course.	CS, IT, EC, EN
10^{th}%	Percentage of marks obtained in 10^{th} class examination.	First $>$ 60%, Second $>$ 45 & $<$ 60%, Third $>$ 35 & $<$ 45%
12^{th}%	Percentage of marks obtained in 12^{th} class examination.	First $>$ 60%, Second $>$ 50 & $<$ 60%
B. Tech	Percentage of marks obtained in B.Tech course.	First $>$ 60%, Second $>$ 50 & $<$ 60%
Final Grade	Final Grade obtained after analysis the passing percentage of 10^{th}, 12^{th}, B.Tech.	Excellent, Good, Average

3.1. Previsão de Desempenho de Alunos

O monitoramento de desempenho envolve avaliações que possuem um papel vital no fornecimento de informações, que é voltado para ajudar os alunos, professores e administradores na tomada de decisões [Pellegrino et al. 2001]. Os fatores de mudança na educação contemporânea têm levado à busca de eficácia e eficiência no monitoramento do desempenho dos alunos em instituições de ensino, que agora está se afastando das técnicas tradicionais de medição e avaliação, para o uso da mineração de dados, que emprega várias técnicas e métodos de investigação para isolar informações importantes que estão implícitas ou ocultas.

O desempenho dos alunos em cursos universitários é de grande importância para o ensino superior, onde vários fatores podem influenciar nos resultados obtidos [Osmanbegović and Suljić 2012]. Em uma universidade, o desempenho geral de um aluno é determinado pelos resultados de avaliações realizadas. Tais avaliações podem ser de diferentes tipos, como desempenho em sala de aula, provas, questionários, trabalhos de laboratório, monitorias, envolvimento em atividades adicionais, currículo, etc.

No objetivo de avaliar e prever o desempenho de estudantes, processos de mineração de dados, especialmente a classificação, vêm sendo aplicados para ajudar no reforço da qualidade do sistema de ensino superior por meio da avaliação dos estudantes. Singh e Kumar [Singh and Kumar 2012] utilizam o método de classificação por árvores de decisão para gerar regras a partir de uma seleção de atributos que podem influenciar o desempenho dos alunos do curso de Bacharelado em Tecnologia do *Bansal Institute of Engineering & Technology*. Estas regras são posteriormente estudadas e avaliadas, e através de um sistema que as utiliza, é possível prever a proporção de estudantes que falham ou são aprovados no exame final.

O primeiro passo para a classificação, neste caso, foi determinar os grupos em que os alunos poderiam se encaixar. Para isso, foram determinados três grupos de acordo com o desempenho: *médio*, com desempenho entre 35 (inclusive) e 45%; *bom*, com desempenho entre 45 (inclusive) e 60%; e *excelente*, com desempenho maior ou igual a 60%. O segundo passo foi determinar os atributos mais significativos, que estão listados na Tabela 1, onde também é feita uma descrição dos mesmos e citados quais valores estes atributos podem assumir. A partir da base de dados dos alunos é montada a árvore de decisão, em que cada nó ramo vai representar uma escolha entre as possíveis alternativas, e cada folha representará uma decisão.

A partir desta árvore de decisão, as regras foram extraídas e aplicadas aos dados de um conjunto de 40 alunos, de quatro "ramos" distintos do curso de Bacharelado em Tecnologia, que compõem a Tabela 2. A Tabela 3 apresenta o resultado da classificação.

Tabela 2. Siglas e nomes das áreas do curso de Bacharelado em Tecnologia

Branch Code	Name of Branches
CS	Computer Science & Engineering
IT	Information Technology
EC	Electronics & Communication Engineering
EN	Electrical & Electronics Engineering

Tabela 3. Classificação final dos estudantes

Branch	No. Students	No. Students Excellent	No. Students Good	No. Students Average
CS	10	6	3	1
IT	10	5	3	2
EC	10	4	4	2
EN	10	5	3	2
Total	40	20	13	7

Objetivando melhorar o desempenho geral de uma instituição, as performances individuais devem ser examinadas. Por isso, é útil para as instituições de ensino analisar o desempenho de seus alunos para identificar as áreas de fraqueza, orientando seus alunos para um futuro melhor. Khatwani e Arya [Khatwani and Arya 2013] propõem um algoritmo baseado em árvores de decisão e algoritmos genéticos, para prever o desempenho de um aluno. O algoritmo ID3 é utilizado para criar várias árvores de decisão, cada uma prevendo o desempenho de um aluno com base em um conjunto de características diferentes. Uma vez que cada árvore de decisão fornece uma visão para o desempenho provável de cada estudante e diferentes árvores dão resultados diferentes, não é possível prever o desempenho, mas é possível identificar as áreas ou características que são responsáveis pelo resultado previsto. Para maior precisão, é incorporado um algoritmo genético que realiza iterações evolutivas para refinar os resultados.

Em [Romero et al. 2013], os autores propõem uma ferramenta de mineração de dados educacionais integrada no ambiente educacional, de maneira que todos os processos de mineração de dados possam ser realizados pelo próprio instrutor, em uma aplicação única, possibilitando que os resultados obtidos sejam aplicados diretamente no ambiente educacional. Para coletar os dados dos usuários, foi desenvolvido um *Moodle* específico, que foi aplicado em diversas universidades, sendo a primeira a Universidade de Córdoba. Os dados coletados são tratados com técnicas de árvores de decisão, redes neurais e lógica *fuzzy*, a fim de prever as notas que os estudantes universitários irão obter no exame final de seu curso.

Um problema recorrente é o da grande quantidade de dados acumulados nas bases do sistema de administração educacional, que são carentes de ferramentas de análise inteligentes. Long e Wu [Long and Wu 2012] utilizaram o algoritmo de árvore de decisão C4.5 [Quinlan 1993] para construir um modelo de análise de desempenho dos alunos que tem como objetivo fornecer as informações necessárias para a melhora da qualidade do ensino e até mesmo para tomada de decisões em futuras reformas no sistema atual.

Na Figura 1, é mostrado o fluxograma do procedimento adotado em [Long and Wu 2012] para o desenvolvimento do modelo de classificação. A primeira

Tabela 4. Atributos e valores que podem ser assumidos

Attribute name	Code	Attribute Value
Total result	A	c1, c2, c3 (excellent, good, poor)
Experimental result	B	c1, c2, c3 (excellent, good, poor)
Paper diff	C	d1, d2, d3 (high, medium, low)
Teacher eval	D	c1, c2, c3 (excellent, good, poor)
Interest	E	y, n (interested, uninterested)

parte realizada no processo de mineração foi a integração dos dados de várias bases, em uma base unificada. Como alguns campos de dados podem ser considerados como atributos de baixa relevância ou redundantes para a classificação do desempenho, o segundo passo é um processo de redução e generalização de atributos, que reflete diretamente na redução da sobrecarga computacional desnecessária.

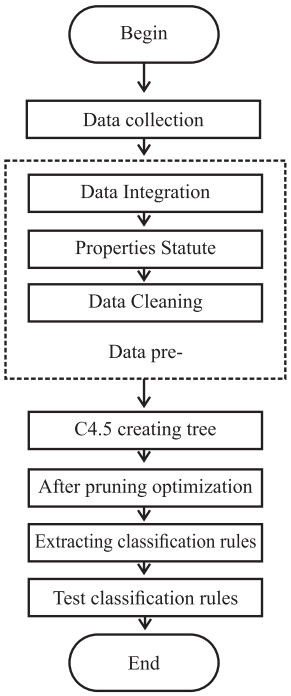

Figura 1. Fluxograma da abordagem proposta em [Long and Wu 2012].

Após a unificação da base de dados e da remoção dos dados que não possuem importância para a classificação, Long e Wu [Long and Wu 2012] aplicam um processo de limpeza dos dados (*data cleaning*) para organizar os dados que foram generalizados. A Tabela 4 mostra os atributos significativos para o problema e os possíveis valores que estes podem assumir.

Na construção da árvore de decisão, os autores empregam o algoritmo C4.5 [Quinlan 1993]. Para a escolha do melhor atributo, são realizados diversos cálculos como a entropia e o ganho de informação. O atributo com maior ganho de informação, por exemplo *B* (Figura 2) se torna o nó raiz da análise de desempenho dos alunos. De acordo com os possíveis valores de *B*, são criados diferentes ramos. O processo é repetido com os demais atributos até gerar a árvore de decisão que é mostrada na Figura 2 (esquerda).

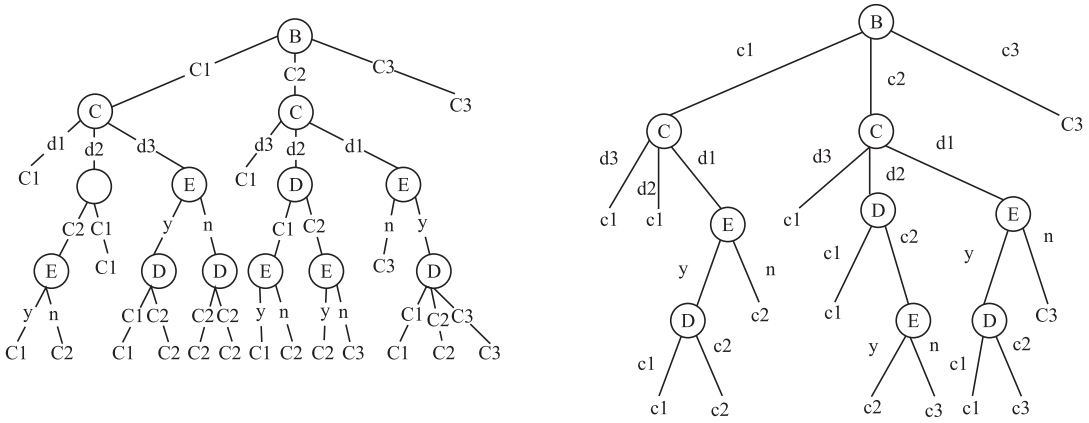

Figura 2. Árvore de decisão gerada a partir dos atributos, antes e depois do processo de poda.

No processo de criação de uma árvore de decisão, por causa dos possíveis ruídos presentes em algumas das amostras, bem como os possíveis valores incorretos no conjunto de treinamento, podem ocorrer anomalias em alguma ramificação da árvore. A fim de melhorar os resultados da árvore gerada e também reduzir sua complexidade, é utilizado um processo de poda. A Figura 2 (direita) resulta da poda da árvore da Figura 2 (esquerda). Após o processo de poda, as regras de classificação são extraídas e utilizadas para classificar as amostras da base de dados.

4. Conclusões

As metodologias de aprendizagem estão ultrapassando os limites dos ambientes das salas de aula tradicionais, contribuindo desta forma para o surgimento de ambientes dinâmicos e informatizados. Uma das consequências oriundas desta migração é o crescimento contínuo dos volumes dos dados armazenados em bases educacionais. Isto, aliado à atual mudança nas metodologias de ensino, tem estimulado o desenvolvimento de uma série de ferramentas para gerar informações e novos conhecimentos a partir do processamento destes dados. Em razão disto, há um crescente interesse na aplicação de técnicas de mineração de dados em ambientes educacionais.

Este trabalho apresentou o levantamento do estado da arte de EDM na previsão de desempenho de alunos. Esta análise possibilitou conhecer melhor as técnicas de mineração utilizadas em trabalhos relevantes e atuais da área. Foi possível constatar que as árvores de decisão, apesar de ser um método de aprendizagem de máquina tradicional, é uma técnica bastante válida na mineração de dados educacionais.

Referências

Aviad, B. and Roy, G. (2011). Classification by clustering decision tree-like classifier based on adjusted clusters. *Expert Systems with Applications*, 38(7):8220–8228.

Baker, R. et al. (2010). Data mining for education. *International Encyclopedia of Education*, 7:112–118.

Castro, F., Vellido, A., Nebot, À., and Mugica, F. (2007). Applying data mining techniques to e-learning problems. In *Evolution of teaching and learning paradigms in intelligent environment*, pages 183–221. Springer.

Connolly, T. M. and Begg, C. E. (2005). *Database systems: a practical approach to design, implementation, and management*. Addison-Wesley Longman.

Han, J., Kamber, M., and Pei, J. (2006). *Data mining: concepts and techniques*. Morgan kaufmann.

Hanna, M. (2004). Data mining in the e-learning domain. *Campus-wide information systems*, 21(1):29–34.

Khatwani, S. and Arya, A. (2013). A novel framework for envisaging a learner's performance using decision trees and genetic algorithm. In *Computer Communication and Informatics (ICCCI), 2013 International Conference on*, pages 1–8. IEEE.

Koedinger, K., Cunningham, K., Skogsholm, A., and Leber, B. (2008). An open repository and analysis tools for fine-grained, longitudinal learner data. *Educational Data Mining*, 1:157–166.

Long, X. and Wu, Y. (2012). Application of decision tree in student achievement evaluation. In *Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on*, volume 2, pages 243–247. IEEE.

Merceron, A. and Yacef, K. (2005). Educational data mining: a case study. *Artificial Intelligence in education: supporting learning through Socially Informed Technology.–IOS Press*, pages 467–474.

Osmanbegović, E. and Suljić, M. (2012). Data mining approach for predicting student performance. *Economic Review*, 10(1).

Pellegrino, J. W., Chudowsky, N., Glaser, R., et al. (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Raghavan, N. S. (2005). Data mining in e-commerce: A survey. *Sadhana*, 30(2-3):275–289.

Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., and Ventura, S. (2013). Web usage mining for predicting final marks of students that use moodle courses. *Computer Applications in Engineering Education*, 21(1):135–146.

Romero, C. and Ventura, S. (2010). Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6):601–618.

Romero, C., Ventura, S., and De Bra, P. (2004). Knowledge discovery with genetic programming for providing feedback to courseware authors. *User Modeling and User-Adapted Interaction*, 14(5):425–464.

Singh, S. and Kumar, V. (2012). Classification of student's data using data mining techniques for training & placement department in tecnical education. In *International Journal of Computer Science and Network (IJCSN)*, pages 121–126.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P., Zhou, Z.-H., Steinbach, M., Hand, D., and Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37.