Mineração de Dados Educacionais para a Predição de Desempenho Acadêmico em Cursos Universitários

Maitê Marques, Maria do Carmo Nicoletti

Faculdade Campo Limpo Paulista - FACCAMP
Campo Limpo Paulista - SP

{maite.marques184@gmail.com, carmo@cc.faccamp.br}

Abstract. *Although originally being business oriented, data mining techniques nowadays can be applied to a variety of types of data such as those from meteorology, medicine, finance, insurance and, particularly, educational environments. This paper presents an ongoing research work in the area of Educational Data Mining, having as its mail goal the prediction of university students' performance in a conventional teaching environment.*

Resumo. *Embora originalmente direcionadas a negócios, presentemente as técnicas de mineração de dados podem ser aplicadas a qualquer tipo de dado, tais como os relacionados à meteorologia, finanças, seguros e, particularmente, aqueles produzidos por instituições educacionais. Este artigo apresenta um trabalho de pesquisa em andamento na área Mineração de Dados Educacionais, cujo principal objetivo é a predição do desempenho de estudantes universitários em ambiente convencional de ensino.*

1. Introdução

A possibilidade de prever o desempenho de alunos em disciplinas de cursos universitários tem papel relevante em ambientes educacionais uma vez que pode subsidiar, de maneira decisiva, a condução do processo ensino-aprendizagem. Antever, com alguma margem de tempo, o desempenho de um aluno pode, entre outros, motivar o docente responsável pela disciplina a usar e aplicar recursos pedagógicos que permitam, por exemplo, a recuperação de alunos que satisfazem a um determinado perfil de desempenho, ainda durante o tempo regular da disciplina, sem a necessidade de tempo extra de recuperação, normalmente requerido. De uma maneira simplista, uma maneira de antever o desempenho escolar é considerar dados relativos a desempenhos escolares anteriores e, com base nesses dados, predizer um desempenho futuro. Obviamente existe um número bem grande de variáveis envolvidas em tal processo e, portanto, dependendo da qualidade dos dados disponíveis, a previsão pode ou não ser satisfatória e, eventualmente, ser usada para corrigir eventuais problemas.

Algoritmos e técnicas da área de Mineração de Dados (MD) podem subsidiar o processo de predição, desde que exista um volume suficiente de dados representativos e confiáveis. A MD pode ser caracterizada, de maneira abrangente e no que tange à pesquisa, como a área voltada ao estudo e investigação de formalismos matemáticos e técnicas computacionais, com vistas ao desenvolvimento de algoritmos para a exploração do volume (cada vez maior) de dados, relativos a um número crescente de domínios de conhecimento. Por meio do uso de algoritmos de MD pretende-se a extração de informação implícita que seja potencialmente útil (tal como regras de associação entre variáveis), a partir de dados relativos a um determinado domínio de conhecimento. Muitas das técnicas caracterizadas como de Mineração são algoritmos da área de Aprendizado de Máquina

(AM) que foram customizados para lidar com grandes volumes de dados e muitas vezes, algoritmos de MD e AM são entendidos como tendo a mesma funcionalidade.

A área intitulada Mineração de Dados Educacionais (MDE) diz respeito ao uso de técnicas de MD em domínios que abrangem dados educacionais. Como apontado em [García *et al.* 2011] e [Baker & Yacef 2010], a MDE emergiu nos últimos anos como uma área de pesquisa para pesquisadores das mais diversas áreas (e.g., Computação, Educação, Estatística, Sistemas Tutores Inteligentes, E-learning, etc.), quando da análise de grandes volumes de dados com o objetivo de resolver questões voltadas às áreas educacionais associadas às respectivas áreas de pesquisa individuais. O aumento no número de repositórios com dados acadêmicos de alunos e, também, a rápida obtenção (embora restrita) de tais dados por meio de *queries* a banco de dados, tem viabilizado, de certa forma, o seu uso em um contexto acadêmico, para extração de informações que podem reverter em replanejamento de disciplinas, redirecionamento de práticas pedagógicas, revisão de ementas, e outros. Processos utilizados em MDE podem ser vistos como conversores daqueles dados acadêmicos brutos que foram obtidos por sistemas educacionais tradicionais (sejam eles automatizados ou não), em informação útil ao sistema educacional como um todo, que pode ser utilizada por desenvolvedores de software, professores, pesquisadores educacionais, etc.

O projeto em desenvolvimento, foco deste artigo, tem por objetivo principal a investigação do uso de determinadas técnicas de MD/AM com o objetivo de prever o desempenho de alunos em disciplinas de cursos universitários. A Seção 2 brevemente apresenta algumas possíveis técnicas de MD cogitadas para a inferência de desempenho acadêmico, com base nos dados descritos na seção seguinte. A Seção 3 descreve os dados que cogitados para os experimentos de MDE, que estão armazenados em um banco de dados de uma instituição educacional localizada no estado de São Paulo. São dados relativos à caracterização de alunos bem como à descrição dos respectivos históricos escolares. A Seção 4 finaliza o trabalho elencando os próximos passos para a conclusão do projeto.

2. Uma Breve Descrição das Técnicas de MD/AM Cogitadas

O projeto tem por objetivo o uso de técnicas de MD/AM para a predição do desempenho de alunos de um curso superior de Bacharelado em Ciências da Computação. Pretende-se também uma comparação entre os resultados obtidos pelas técnicas cogitadas com o objetivo não apenas de descobrir qual é a mais apropriada para o objetivo proposto mas, também, com vistas a um futuro sistema computacional que a disponibilize a docentes, como parte de uma ferramenta adicional para acompanhamento de alunos. Aprendizado indutivo de máquina é um processo de aprendizado automático que, a partir de um conjunto de instâncias (exemplos, padrões) induz uma expressão geral do conceito representado pelas instâncias fornecidas. Caso as instâncias sejam descritas também por um atributo classe e o algoritmo faça uso dessa informação, o aprendizado é chamado de supervisionado, caso contrário, de não supervisionado. Dependendo do algoritmo utilizado a expressão geral pode ser representada como um conjunto de regras de decisão, uma árvore de decisão, uma rede neural, ou simplesmente como um conjunto de grupos formados com base nas instâncias fornecidas. Para o objetivo de aprendizado proposto, em princípio, serão utilizados: (1) classificadores Naïve Bayes [Friedman *et al.* 1997]; (2) árvores de decisão (algoritmo C4.5) [Quinlan 1993]; (3) o NN [Cover & Hart 1967] e o (4) Apriori [Agrawal e Srikant 1994] que são brevemente descritos a seguir.

Classificadores Naïve Bayes (CNB) são uma família de classificadores probabilísticos simples, baseados no uso do resultado estabelecido pelo teorema de Bayes. O CNB assume que: (i) todas as variáveis (atributos) são condicionalmente independentes umas das outras, dada a variável classe e (ii) todas as variáveis são diretamente dependentes da variável classe. Langley e colegas em [Langley *et al.* 1992] mostraram que o CNB consegue competir com um dos mais bem sucedidos sistemas de aprendizado de máquina, conhecido como C4.5 [Quinlan 1993]. O C4.5 constrói árvores de decisão a partir de um conjunto de dados de treinamento, utilizando para direcionar o crescimento da árvore, o conceito de entropia da informação. O método Nearest Neighbor (NN) é um método de aprendizado automático caracterizado como *baseado em instâncias*; é bastante popular para a classificação de instâncias de dados devido, principalmente, à sua simplicidade e habilidade de produzir classificações com boa precisão. O algoritmo conhecido como Apriori é voltado à mineração de padrões sequenciais. No trabalho em que foi proposto [Agrawal & Srikant 1994] o algoritmo descrito gera, como saída, um conjunto de itemsets considerados frequentes (um itemset é um conjunto de itens básicos). A partir do conjunto de itemsets frequentes, são geradas regras de associação, com o objetivo de estabelecer possíveis relações entre itens de dados participantes de *itemsets*, para uso futuro, geralmente por sistemas administrativos de tomada de decisão. A extração de regras de associação a partir de dados educacionais já foi objeto de algumas pesquisas descritas na literatura (ver [García *et al.* 2011], [Abdullah *et al.* 2011]) com relativo sucesso.

3. Sobre os Dados a serem Utilizados

Como apontado em [Whitten & Frank 2005], técnicas de mineração de dados investem no desenvolvimento de programas computacionais que automaticamente "peneiram" bases de dados, na busca de regularidades ou padrões. Tais padrões, quando robustos, podem ser generalizados e usados em predições relacionadas a dados futuros. É importante lembrar que, muitas vezes, processos de mineração encontram padrões que são triviais ou de senso comum. Outras vezes, espúrios ou incertos. Um aspecto de extrema relevância quando do uso de técnicas de aprendizado de máquina e/ou mineração de dados, diz respeito à qualidade dos dados disponíveis para a tarefa de aprendizado/mineração. Via de regra dados reais (i.e., dados que não são artificialmente criados para validar certos aspectos de algum sistema) têm imperfeições: podem ter partes confusas bem como ausentes e, como consequência, qualquer padrão detectado provavelmente será inexato. Os autores enfatizam que algoritmos precisam ser robustos o suficiente para lidar com dados imperfeitos e para extraírem regularidades que, embora inexatas, sejam úteis.

Os dados a serem utilizados nos experimentos com técnicas de MDE provêm de um curso superior de Bacharelado em Ciências da Computação oferecido por uma faculdade da região metropolitana da cidade de São Paulo que, por razões de sigilo será referenciada neste trabalho apenas como Faculdade. O curso de Bacharelado em Ciência da Computação foi escolhido devido à maior familiaridade com a grade curricular das autoras deste trabalho. Alunos ingressam no curso via exame de seleção promovido pela Faculdade, duas vezes ao ano. Para alunos que seguem a matriz curricular vigente, a duração do curso é de 4 anos. Via de regra a população estudantil provêm de regiões relativamente próximas à Faculdade, sendo um pequeno número proveniente de outros estados e alguns poucos estrangeiros. Informações sobre alunos da Faculdade são armazenadas em dois

tipos de registros: o Registro do Estudante e o Histórico Escolar do Estudante. As informações contidas em cada um deles estão na Tabela 1. Nos experimentos do Registro do Estudante serão utilizados inicialmente as informações (5), (6), (7), (8) e (12) e do Registro de Dados Acadêmicos (4), (9) e (10). O impacto de cada um dos atributos nos resultados será avaliado empiricamente, quando da realização dos experimentos. Estima-se que o número de padrões de dados para os experimentos, vai estar em torno de 240. Com base em registros históricos de desempenho no curso e, particularmente, na disciplina de Construção de Algoritmos e Programação I (CAP1), serão induzidos classificadores, para predição do desempenho na disciplina CAP1.

Tabela 1. Informações Discentes junto à Faculdade.

REGISTRO DO ESTUDANTE	COMENTÁRIOS E/OU OPÇÕES DISPONÍVEIS
(1) Nome	
(2) Endereço	
(3) Telefone	
(4) E_mail	
(5) Data de nascimento	
(6) Idade	
(7) Sexo	
(8) Estado Civil	solteiro(a), casado(a), divorciado(a), viúvo(a).
(9) RG	
(10) CPF	
(9) Título de eleitor	
(9) RNE	para estrangeiros apenas.
(10) Cor	amarela, branca, indígena, negra, parda, não declarado.
(11) Religião	61 possíveis religiões disponíveis para escolha.
(12) Maior grau de instrução	médio completo, superior incompleto, superior completo, especialista, mestre, doutor.
(13) Nome do pai	
(14) Nome da mãe	
(15) Nome do responsável	
DADOS ACADÊMICOS	**COMENTÁRIOS E/OU OPÇÕES DISPONÍVEIS**
(1) Registro Acadêmico (RA)	
(2) Data de Conclusão Ensino Médio	
(3) Data do vestibular	
(4) Nota no vestibular	
(5) Classificação no vestibular	
(6) Data ingresso no curso	
(7) Data término do curso	
(8) Turma	Reserva de vaga, cursando, concluído.
(9) Lista de Disciplinas Cursadas	Cada disciplina cursada tem as seguintes informações a ela associadas: [Nome,Turma,Sem.,Ano,NroFaltas,Notas,Média].
(10) Histórico escolar	Aproveitamentos, disciplinas cursadas, disciplinas a cursar, pendências.

4. Próximos Passos do Desenvolvimento do Projeto

A pesquisa descrita se iniciou com um levantamento bibliográfico preliminar, com o objetivo de prospectar a área de MDE e identificar algumas das contribuições existentes. Em seguida foi feito um estudo de técnicas de MD/AM, com vistas a selecionar as mais promissoras para cumprir o objetivo pretendido da pesquisa que, presentemente, se encontra na fase definição e pré-processamento dos dados advindos de um ambiente educacional específico, de maneira a torná-los factíveis de serem tratados por técnicas de

MD/AM. Particularmente estão sendo definidas e realizadas tarefas relacionadas à seleção e transformação de atributos e, eventualmente, de integração de dados. Uma vez finalizada a fase atual, as técnicas de AM/MD brevemente descritas na Seção 2 serão usadas em experimentos relacionados à predição de desempenho acadêmico, com análise e discussão dos resultados obtidos, bem como recomendações quanto à efetividade do uso de tais técnicas como uma ferramenta adicional ao acompanhamento do desempenho discente. Para os experimentos pretende-se usar o ambiente Weka de MD/AM [Weka 2012].

Referências

[Abdullah *et al.* 2011] Abdullah, Z.; Herawan, T.; Ahmad, N.; Deris, M. M. (2011) Mining significant association rules from educational data using critical relative support approach, Procedia - Social and Behavioural Sciences, v. 28, 97−101.

[Agrawal & Srikant 1994] Agrawal, R.; Srikant, R. (1994) Fast algorithms for mining association rules in large databases, In: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), Chile, 487−499.

[Baker & Yacef 1010] Baker, R.; Yacef, K. (2010). The state of educational data mining in 2009: A review and future visions, Journal of Educational Data Mining, 3−17.

[Cover & Hart 1967] Cover, T. M.; Hart, P. E. (1967) Nearest neighbor pattern classification, IEEE Transactions on Information Theory, v. IT-13, no. 1, 21–27.

[Friedman *et al.* 1997] Friedman, N.; Geiger, D.; Goldszmidt, M. (1997) Bayesian network classifiers, Machine Learning, v. 29, 131−163.

[García *et al.* 2011] García, E.; Romero, C.; Ventura, S.; de Castro, C. (2011) A collaborative educational association rule mining tool, Internet and Higher Education, v. 14, 77−88.

[Langley *et al.* 1992] Langley, P.; Iba, W.; Thompson, K. (1992) An analysis of Bayesian classifiers, In: Proc. of the AAAI-92, 223-228.

[Mitchell 1997] Mitchell, T. M. (1997) *Machine Learning*, USA: McGraw-Hill.

[Quinlan 1993] Quinlan, J. R. (1993) Programs for Machine Learning. Morgan Kaufmann Publishers, Inc. USA:

[Theodoridis & Koutroumbas 2009] Theodoridis, S.; Koutroumbas, K. (2009) *Pattern Recognition*, 4th ed., USA: Elsevier.

[Weka 2012] *WEKA Manual*, Version 3-6-8, New Zealand: University of Waikato.

[Whitten & Frank 2005] Whitten, I. H.; Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd. ed., USA: Elsevier.