Avaliação do Perfil de Uso no Ambiente Moodle Utilizando Técnicas de Mineração de Dados

Leandro C. Santana[1], Alexandre M. A. Maciel [1], Rodrigo L. Rodrigues[2]

[1] Universidade de Pernambuco (UPE)
Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

[2] Universidade Federal Rural de Pernambuco (UFRPE)
Departamento de Educação (DEd)

{lcs, amam}@ecomp.poli.br, rlr@ded.ufrpe.br

Abstract. *Virtual learning environments make the extraction of variables that can be used for data mining increasingly possible. Extract relevant information that will enable the effective monitoring of students in courses mediated by technology is a challenge. This article proposes the application of classification techniques on a set of learning from a course in blended mode data, in order to obtain results as a way of supporting decision made by teachers and administrators. For the development of this work, the profile of Use AVA interaction was considered, in order to make possible classifications, taking into account variable as target student achievement. As satisfactory results, this work obtained performance rate of 74% accuracy, applying the technique of decision tree J48.*

Resumo. *Os Ambientes virtuais de aprendizagem possibilitam cada vez mais, a extração de variáveis que podem ser utilizadas para mineração de dados. Extrair informações relevantes que viabilizem o acompanhamento efetivo de estudantes em cursos mediados por tecnologia é um desafio. Este artigo teve como proposta a aplicação de técnicas de classificação em um conjunto de dados educacionais a partir de um curso na modalidade semipresencial, objetivando a obtenção de resultados como forma de apoiar a tomada de decisão por parte de professores e gestores. Para o desenvolvimento do trabalho foi considerada a interação Perfil de Uso do AVA, a fim de fazer possíveis classificações, levando em consideração como variável alvo o desempenho do aluno. Como resultados satisfatórios obtivemos a aplicação da técnica de árvore de decisão J48, tendo como performance uma taxa de 74% de acurácia.*

1. Introdução

Nos últimos anos os sistemas de ensino baseados em web cresceram exponencialmente, estimulado pelo fato que tanto o professor quanto os alunos não têm a obrigação de um local especifico e que essa forma de ensino independe de plataforma computacional ou hardware específico (Brusilovsky and Peylo, 2003).

Segundo Mostow e Beck (2005) esses tipos de sistemas educacionais podem acumular grande quantidade de informação que são extremamente valiosas para a análise do comportamento dos alunos. Informações tais como a leitura, a escrita, exercícios, provas e até a comunicação com seus pares podem ser gravadas em logs e facilmente acessadas.

Devido a essa vasta quantidade de dados que esses sistemas podem gerar diariamente, é muito difícil gerenciá-las manualmente, e a demanda por ferramentas que possam auxiliar nessa tarefa é grande.

Nesse sentido, surgiu em 2008 a primeira conferência internacional em Mineração de Dados Educacionais (no inglês: Educational Data Mining – EDM). Para Ryan (2011) o principal foco da EDM é o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais, a fim de compreender de forma mais eficaz os alunos, e outros fatores que influenciam a aprendizagem. Esse novo conhecimento pode ser explorado não só pelos alunos, mas também pelos professores/tutores, que passa dessa forma orientar de forma eficiente, levando em consideração o ponto de vista de um conjunto ou de um determinado aluno (Zorrilla et al., 2005).

As variáveis que podem ser mineradas dentro de uma AVA são inúmeras e podem rotuladas a partir de características ligadas a um determinado conjunto de informações. Em Moore (1989) foi apresentada a teoria da interação, que visa distribuir as informações encontradas em um ambiente educacional em um subgrupo de dados, na qual chamou de dimensões, que são elas: Perfil de Uso, Interação Estudante-Estudante e Interação Bidirecional Estudante-Professor.

Em Gottardo (2014) foram utilizadas as dimensões propostas por Moore (1989) e realizados experimentos para estimar o desempenho de estudantes por meio destas dimensões utilizando classificadores. Apesar dos bons resultados, alguns atributos importantes deixaram de ser avaliadas, como: número total de acesso ao fórum, número total de interações com as vídeo-aulas, número total de interações com o livro digital da disciplina, número total de interações com as apresentações em Slides e o tempo médio de acesso no ambiente.

Este trabalho tem como objetivo realizar a avaliação da dimensão perfil de uso no ambiente Moodle utilizando técnicas de mineração de dados, especificamente técnicas de classificação de padrões. O restante do artigo está organizado da seguinte forma: A seção 2 apresenta a revisão da literatura referente ao foco deste trabalho, a seção 3 apresenta os experimentos realizados, a seção 4 os resultados obtidos e a discussões, e por último a seção 5 apresenta as considerações finais e as possibilidades de trabalhos futuros.

2. Revisão da Literatura

Nesta seção iniciamos a revisão da literatura na qual são apresentados os conceitos necessários à fundamentação deste trabalho. Entre os conceitos a serem explorados estão: Teoria da Interação e Mineração de Dados Educacionais.

2.1 Teoria da Interação

A Teoria da Interação foi proposta por Moore em 1989 no propósito de responder perguntas acerca de uma problemática levantada em uma reunião realizada em Salt Lake City em 16 de Abril de 1989. Neste evento foram debatidas questões como: Qual o nível de interação é essencial para uma aprendizagem eficaz? O que é podemos dizer se é bom ou não e uma interação? E como podemos alcançá-lo? Interação em tempo real contribui em algo? O custo disso é levado em consideração?

Diante dessa problemática, Moore sugeriu que educadores à distância precisariam concordar que com o surgimento de três tipos de interação, que são intitulados de Interação ou Dimensão Aluno-Conteúdo, Aluno-Professor e Estudante-Estudante. Para distinguir entre estes três tipos, abaixo teremos uma definição de cada dimensão.

A primeira interação é entre o aluno e o conteúdo ou tema de estudo. Sem ela, não pode haver educação, uma vez que é o processo de interação com o conteúdo intelectual que resulta em mudanças na compreensão do aluno, e perspectiva do mesmo. Esse tipo de interação, Holmberg (1986) chama de "conversa didática interna" quando os alunos "falam para si mesmos", sobre as informações e ideias que eles encontram em um texto, programa de televisão, palestra, ou em outro lugar.

O segundo tipo de interação considerado como essencial por muitos educadores, é a interação entre o aluno e o professor. Nessa interação, os professores procuram estimular ou, pelo menos, manter o interesse do aluno no que deve ser ensinado, para motivar o aluno a aprender.

E por fim, a terceira dimensão da educação à distância, um desafio para o nosso pensamento e prática na década de 1990. A interação entre um aluno e outros alunos, individualmente ou em grupo, com ou sem a presença em tempo real de um professor. Phillips, Santoro e Kuehn (1988) descrevem a importância da interação entre os membros de uma classe que tiveram que aprender habilidades de interação do grupo. Os pesquisadores descobriram que não podiam efetivamente facilitar a interação entre os membros de uma grande classe de graduação em face-a-face salas de aula, e se virou para afastar técnicas de educação, o uso de vídeo gravado e interação com o computador para alcançar um maior desempenho em comportamentos de grupo do que tinham sido capazes de obter em grupos ao vivo.

2.2 Mineração de Dados Educacionais

Após existir alguns eventos relacionados a área, foi apenas em 2005, em Pittsburgh, EUA, que foi organizado o primeiro Workshop on Educational Data Mining. A partir desta data houveram mais eventos relacionados e em 2008 lança-se, em Montreal,

Canadá, a primeira conferência em EDM: First International Conference on Educational Data Mining, evento que ganhou regularidade de realização anual, em sua sétima edição neste ano de 2014. Foi no ano de 2009, que a sociedade investiu na criação de um periódico e publicou o seu primeiro volume do JEDM – Journal of Educational Data Mining. Diante desse cenário, a área de EDM está bem consolidada internacionalmente, porém no Brasil está em desenvolvimento, buscando recursos a fim de consolidar e promover mais eventos relacionados, no intuito de gerar mais pesquisas na área.

Para Baker (2011) em grande parte, as técnicas utilizadas na área são as mesmas de mineração de dados, porém na maioria das vezes é necessário adaptá-las devido às particularidades dos projetos e dos dados. As técnicas podem ser agrupadas diante da taxonomia proposta por Baker et al. (Baker, 2011), como pode ser visto a seguir:

- Predição
    - Classificação
    - Regressão
- Agrupamento
- Mineração de Relações
    - Mineração de Regras de Associação
    - Mineração de Correlações
    - Mineração de Padrões Sequenciais
    - Mineração de Causas
- Destilação de dados para facilitar decisões humanas
- Descobertas com modelos.

As próximas seções descreve a metodologia empregada neste trabalho utiliza as técnicas de classificação de padrões visto que esse tipo de técnica possibilita a previsão de desempenho de alunos de forma bastante consolidada na literatura.

3. Experimentos Realizados

Para a realização deste trabalho foi inspirado na metodologia *CRoss Industry Standard Process for Data Mining* (CRISP-DM), que segue um roteiro de projeto que vai de entender o negócio até sua aplicabilidade, mostrando os modelos e técnicas a ser aplicadas.

3.1 Entendimento dos Dados

Para iniciar uma análise foi necessário fazer um levantamento de atributos que foram utilizados de um banco de dados Moodle. Para a seleção dos atributos utilizamos as indicações de Moore (1989) onde afirma que uma das formas utilizadas para definir quais atributos selecionar e representar os estudantes é basear-se na Teoria de Interação em EAD. Teoria que destaca três interações importantes: entre o estudante e o conteúdo ou objeto de estudo, entre o estudante e o especialista que elaborou o material em questão ou algum outro especialista atuando como instrutor e entre o estudante e outros estudantes, sozinho ou em grupo, com ou sem a presença em tempo real de um instrutor.

Diante das três dimensões para representar os estudantes no AVA, este artigo fez simulações com base apenas na dimensão “perfil de uso do AVA” que tem como

objetivo identificar dados que representem aspectos de planejamento, organização e gestão do tempo do estudante para a realização do curso. Abaixo, na Tabela 1 está descrito todos os atributos utilizados neste trabalho.

Tabela 1: Atributos selecionados para representação de estudantes em um AVA

Dimensão	Atributo	Representação
Perfil de Uso do AVA	Desempenho Final	Result_final
	Número total de acesso ao fórum	Sum_int_forum
	Número total de interações com as vídeo-aulas	Sum_int_video
	Número total de interações com o material da disciplina (Caderno)	Sum_int_mat
	Número total de interações com as apresentações em Slides.	Sum_int_ppt
	Tempo médio de acesso no ambiente	Media_acesso

Para esses experimentos, foram desconsiderados os alunos que desistentes, pois nesses casos os resultados não influenciam. Foi selecionado uma população de 79 estudantes, ilustrado na Tabela 1. É importante salientar que foi feito o processo de discretização no atributo da Tabela 1 "result_final", assim a classe objetivo representa o processo de classificação.

A realização desse experimento se deu a partir de uma base de dados de uma disciplina intitulada Jogos Educacionais, do curso de Licenciatura em Computação no segundo semestre de 2013, realizada pela Universidade Federal Rural de Pernambuco, que contou com 98 alunos do sétimo período na modalidade EAD, dentre os quais apenas 79 alunos concluíram a disciplinas e estes foram os selecionados para o experimento.

3.2 Preparação dos Dados

Foi utilizada a ferramenta Weka para simular e realizar dois experimentos, onde no experimento 1, foi feito o processo de discretização e inserido três classes, que são: A e C para os alunos com as maiores e menores notas no ambiente respectivamente, e conceito B para os demais. No experimento 2 novamente foi feito o processo de discretização e incluído duas classes, que são: Classe "Aprovado" para os alunos com nota igual ou maior que 70, e Classe "Reprovado" com notas inferiores a 70. As tabelas 2 e 3 mostram a distribuição dos dados para as duas discretizações realizadas.

Tabela 2: Distribuição após processo de discretização do experimento

Título da Classe	Descrição	Número de Estudantes	Intervalo de Notas
A	Alunos com notas superiores	11	Maior ou igual a 80
B	Alunos com notas intermediárias	34	Entre 60 e 79
C	Alunos com notas inferiores	34	Menor que 60

Tabela 3: Distribuição após processo de discretização do experimento 2

Título da Classe	Descrição	Número de Estudantes	Notas
Aprovado	Alunos Aprovados	26	Maior ou igual a 70
Reprovado	Alunos Reprovados	53	Menor que 70

3.3 Modelagem e Avaliação

Para o desenvolvimento deste trabalho foram utilizados sete algoritmos de classificação, que são eles: *RandomForest*, *MultilayerPerceptron (MLP)*, *NaiveBayes*, SVM, KNN, J48 e RBF. Para auxiliar na avalição dos resultados, foi utilizada a técnica matriz de confusão, assim os resultados são apresentados em uma matriz bidimensional.

Os elementos da matriz mostram quais elementos foram classificados corretamente ou não. Uma outra avaliação foi a acurácia, que mediu a taxa de acerto global, ou seja faz uma divisão das classificações corretas pelo número total de instâncias dos dados a serem classificados. O método de cálculo da acurácia é conhecido como K - *fold Cross-Validation* que é uma técnica para a estratificação da base dados em um conjunto de treinamento e outro de teste. Estudos relatados sugere-se a adoção do número 10 como valor padrão para o número de partições dos dados K (Witten, 2011).

4. Resultados e Discussões

No experimento 1, foi utilizado todos os atributos da Tabela 1, apenas considerando o processo de discretização da Tabela 2. Ou seja, foi incluída para análise, um novo parâmetro, chamado de "classe". Onde inicialmente foi divido em três atributos, denominados A, B e C, e distribuído as quantidades de alunos que se enquadravam em cada atributo. Para esse experimento, o algoritmo J48 apresentou melhor classificação, com 53,16% taxa de acerto, e a KNN com a classificação mais baixa, 44,30%.

A Tabela 4 mostra as matrizes de confusão do processo de classificação dos dois algoritmos em destaque nesse experimento. Onde a diagonal é referente a classificação correta das instancias, onde está destacada com amarelo.

Tabela 4: Matriz de Confusão dos algoritmos J48 e KNN

J48				KNN			
Correlatas/Previstas	**A**	**B**	**C**	**Correlatas/Previstas**	**A**	**B**	**C**
A	3	5	3	**A**	3	5	3
B	4	10	20	**B**	4	11	19
C	1	4	29	**C**	3	10	21

No experimento 2, foi considerando o processo de discretização da Tabela 3, nessa discretização foi inserida a classe com os atributos Aprovado e Reprovado. E como resultado o algoritmo J48 apresentou melhor classificação, com 74,68% taxa de acerto, e a MLP com a classificação mais baixa, 62,02%. A Tabela 5 mostra as matrizes de confusão dos dois algoritmos.

Tabela 5: Matriz de Confusão dos algoritmos J48 e MLP

J48			MLP		
Correlatas/Previstas	**Aprovado**	**Reprovado**	**Correlatas/Previstas**	**Aprovado**	**Reprovado**
Aprovado	13	13	**Aprovado**	9	17
Reprovado	8	45	**Reprovado**	10	43

Os demais algoritmos do experimento, não foram levados em consideração por possuir uma taxa mediana em relação aos que tiveram destaque nas tabelas citadas acima. Porém, pode ser visualizado os resultados encontrados de todos os algoritmos na Tabela 6.

Tabela 6: Taxa de Acerto dos algoritmos em relação aos dois experimentos

Corr/Pr	RF	MLP	NB	SVM	KNN	J48	RBF
ABC	49,36%	45,56%	53,16%	51,89%	44,30%	53,16%	45,56%
Ap/Rep	65,82%	62,02%	73,41%	73,41%	64,55%	74,68%	72,15%

Foi utilizada a técnica de teste estatístico T pareado no ambiente Weka Experiment Environment – WEE a fim de testar a significância estatística dos resultados obtidos, com nível de significância de 5%. Foi constatado que dentre todas as técnicas utilizadas, a NaiveBayes (NB) e a SVM possuem diferenças diante das amostras estudadas, mostrando não ser flexível quanto aos atributos e classes propostas.

Com base em nossos resultados, percebe-se que as taxas de acerto estão entre 44,30% e 53,16% para o experimento 1, fora de um intervalo aceitável considerando o trabalho de Hämäläinen e Vinni (2011), onde foi possível obter uma taxa de 72%, com base nas pesquisas que envolvem estimativas de desempenho dos perfis de uso do Moodle. Para o experimento 2, quatro algoritmos merecem destaque, pois ultrapassaram a taxa de 72% estimada para desempenho dos perfis de uso.

4.1 Discussões

Ao final deste trabalho diversos pontos importantes foram passíveis de discussão. Um dos pontos mais importantes a destacar refere-se à análise de variáveis utilizadas na EAD, que pertencem a dimensão Perfil de Uso do AVA, proposta por Gottardo (2014). Nesta dimensão pudemos atribuir as variáveis tabela 1. Essas variáveis não estavam presentes no estudo de Gottardo (2014) desse modo torna este trabalho como uma análise complementar aos estudos mencionados.

Outro ponto que mereceu destaque foi a utilização de variáveis que demonstraram a quantidade de acesso a materiais que puderam ser baixados para o computador dos alunos, ou seja, materiais como textos em .pdf podem ser baixados pelos alunos uma única vez e ser utilizado diversas vezes no modo off-line, não possibilitando a captura do log diário pelo ambiente. Esta situação pode ter ocasionado a diminuição das taxas de acurácia dos algoritmos de classificação.

Com os resultados encontrados neste trabalho foi possível a obtenção de limiares que podem ser implementados em ambientes virtuais de aprendizagem para serem utilizados em diferentes cursos para a previsão de índice de aprovação dos alunos. Estes limiares podem ser melhorados a medida que a base de dados de treinamento for aumentando.

5. Conclusão e Trabalhos Futuros

Este trabalho teve a proposta de avaliar um conjunto de atributos do perfil de uso do ambiente Moodle, que não foi abordado no experimento realizado por Gottardo (2014). Os resultados obtidos foram considerados satisfatórios, atingindo taxas de 72% de acurácia na aplicação de determinadas técnicas de classificação.

Os resultados obtidos com o experimento confirmam a viabilidade de se classificadores para a predição de desempenho de estudantes. Isso indica que a proporção da variação da variável desempenho, pode ser inferida em função das variáveis da dimensão *Perfil de uso do AVA*.

Como trabalhos futuros pretende-se realizar um novo experimento com objetivo de averiguar melhorias nas taxas de acurácia. Pretende-se também avaliar se a inclusão de novos atributos pode proporcionar resultados ainda melhores.

Referências

Baker, R.S.J.d., I. S. d. C. A. (2011). Mineração de dados educacionais: Jornada de Atualização em Informática na Educação - JAIE 2012 25/29. Oportunidades para o Brasil. Revista Brasileira de Informática na Educação, 19(2).

Brusilovsky, P., Peylo, C., (2003). Adaptive and intelligent web-based educational systems. International Journal of Artificial Intelligence in Education, 13, 156-169.

Gottardo E., Kaestner C. A. A., Noronha R. V. (2014). Estimativa de Desempenho Acadêmico de Estudantes: Análise da Aplicação de Técnicas de Mineração de Dados em Cursos a Distância. Pág 45.

Hämäläinen, W., Vinni, M. (2011). Classifiers for Educa- tional Data Mining. In: Romero et al. Handbook of Educational Data Mining. Flórida, CRC Press, p. 57-71.

Holmberg, B. (1986). Growth and Structure of Distance Education. London: Croom-Helm.

Moore M. G. (1989). Three Types of Interaction. The American Journal of Distance Education, 3(2):1–6.

Mostow, J., Beck, J., Cen, H., Cuneo, A., Gouvea, E., Heiner, C., (2005). An educational data mining tool to browse tutor-student interactions: Time will tell! In Proceedings of the Workshop on Educational Data Mining, Pittsburgh, USA (pp. 15–22).

Phillips, G. M., G. M. Santoro, and S. A. Kuehn. (1988). The use of computer-mediated communication in training students in group problem-solving and decision-making techniques. The American Journal of Distance Education 2(1):38-51.

Ryan B., Seiji I., Adriana C. (2011). Mineração de Dados Educacionais: Oportunidades para o Brasil. Revista Brasileira de Informática na Educação. Pág. 3.

Zorrilla, M. E., Menasalvas, E., Marin, D., Mora, E., Segovia, J., (2005). Web usage mining project for improving web-based learning sites. In Web Mining Workshop. Cataluna (pp. 1-22).

Witten, I.H., Frank, E., Hall, M.A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann, 3 ed.