Aplicação de mineração de dados na predição de evasão de alunos da educação a distância

Emanuel Marques Queiroga[1]; Cristian Cechinel[2];
[1]*Universidade Federal de Pelotas - UFPEL – emanuelmqueiroga@gmail.com 1*
[2]*Universidade Federal de Pelotas - UFPEL – contato@cristiancechinel.pro.br*

1. INTRODUÇÃO

Em um contexto social onde tornou-se indispensável a busca por conhecimentos e qualificação das pessoas, de forma que nos últimos anos o governo brasileiro através do Ministério da Educação e entidades de fomento tem feito uma série de investimentos em programas de educação buscando a qualificação da mão de obra produtiva no país. Considerando que o Brasil é um país de grandes dimensões, diversas cidades estão afastadas dos grandes centros universitários e acabam ficando isoladas de programas de graduação e cursos técnicos profissionalizantes. Desta forma, uma das alternativas adotadas pelo governo federal para a expansão do acesso a educação foi a utilização da modalidade à distância (Educação a Distância - EAD), que tem como um de seus objetivos levar o ensino a estas localidades, geralmente utilizando Ambientes Virtuais de Aprendizagem (AVA) (DELANO, 2013).

Entretanto, um dos principais desafios da EAD é obter a diminuição do índice de evasão, que conforme o Censo EAD (CensoEAD, 2013), foi de 18,6% em 2010, 20,5% em 2011, 11,74% em 2012 e 16,94% em 2013 nos cursos autorizados pelo Ministério da Edu-cação (MEC). Num contexto onde em 2013 haviam 5.754 cursos autorizados pelo MEC e a taxa de matriculas anual foi de 882.843, temos em torno de 149.553 alunos evadidos.

Segundo MANHAES (2010), a detecção precoce de grupos de alunos com risco de evasão é uma condição importante para reduzir o problema da evasão, uma vez que um tratamento mais adequado pode ser ofertado a esses alunos. Ainda segundo Manhães, atualmente o processo de identificação desse grupo de alunos é manual, subjetivo, empírico e sujeito a falhas, pois depende primordialmente da experiência acadêmica e do envolvimento dos docentes.

Assim a aplicação da mineração de dados, que é a disciplina que estuda a descoberta de novas informações a partir da analise de grandes quantidades de dados e tem como objetivo identificar relações e padrões nos dados e assim produzir novas informações, pode ajudar na busca pela redução no número de alunos evadidos.

A mineração de dados educacionais (EDM) é uma área de pesquisa recente e que tem como principal objetivo o desenvolvimento e aplicação de técnicas de mineração de dados na exploração de conjuntos de dados coletados em ambientes educacionais. Atualmente a EDM vem se estabelecendo como uma forte e consolidada linha de pesquisa que possui grande potencial para melhorar a qualidade do ensino (BAKER, 2011). Essa área pode ajudar as instituições a criarem modelos de predição que tenham condições de avaliar as chances de um determinado acadêmico evadir.

2. METODOLOGIA

Para o desenvolvimento deste projeto foram utilizados dados de dois cursos técnicos na modalidade a distância do Instituto Federal Sul-riograndense (IFSul). Estes cursos funcionam com atividades semanais que são postadas no ambiente pelo professor e os alunos tem uma semana para o desenvolvimento desta com auxilio dos tutores. Cada curso tem um tempo de realização máximo de 103 semanas dentro de 24 e a situação final do aluno é determinada pelo seu resultado nas avaliações. O aluno assim pode assumir 2 estados diferentes no final das atividades, aprovado ou reprovado, entretanto este estudo tem como objetivo a predição dos alunos que se evadiram no decorrer do curso. Para tal definisse que o aluno será considerado evadido caso abandone não efetuando as atividades no decorrer do curso e não efetue sua matriculara no semestre seguinte.

Como este trabalho propõe modelos de fácil generalização e que assim possam ser aplicados em outros cursos do IFSul ou até mesmo em outras instituições de ensino, se optou por utilizar somente a contagem semanal de interações dos alunos. Para isto a metodologia foi dividida nos seguintes passos: coleta de dados, pré-processamento dos dados e geração e avaliação de modelos de predição.

Na coleta de dados foram coletados 1.716.683 log's de interações dos alunos, para o pré-processamento deste volume considerável de dados foi desenvolvido um sistema em Java que tem como objetivo automatizar esta e a próxima etapa do projeto. Este sistema efetua os calculos das atividades a partir das datas de inicio e fim do semestre que são incluidas nele e salva os valores em um banco de dados.

Para a etapa de geração dos modelos de predição foi utilizada a biblioteca de desenvolvimento do WEKA compatível com Java e integrada diretamente ao sistema desenvolvido. Foram definidos 3 tipos de experimentos diferentes, no primeiro se treina e aplica o modelo nos dados do curso 1, no segundo se treina e aplica o modelo nos dados do curso 2 e no terceiro se treina com um curso e o modelo gerado é aplicado no outro curso. Os algoritmos utilizados foram, Bayes Net (BN), Simple Logistic (SL), Multilayer Perceptron (MP), Random Forest (RF) e J48 (J48).

Como o objetivo é a predição precoce dos alunos em risco de evasão são gerados um modelo por semana de curso, então temos 103 modelos por curso onde no primeiro modelo temos os dados somente da semana 1, no segundo das semanas 1 e 2 e assim até o final.

3. RESULTADOS E DISCUSSÃO

Os experimentos obtiveram uma acurácia geral e de verdadeiros negativos muito próxima nos 3 experimentos a partir da vigésima semana. Isso se deve ao fato de que os conjuntos de dados são relativamente balanceados como dito anteriormente.

Nos 3 experimentos os resultados tem um gradativo crescimento com o passar do tempo, tendo os algoritmos mantido um crescimento praticamente regular até atingir seu ápice por volta da semana 70 e continuado estáveis até o final. Apesar disto, alguns dos algoritmos não apresentaram nenhum nível de

aprendizagem nas primeiras semanas como é o caso do Bayes Net no segundo experimento e do Simple Logistic, Redes Neurais e Bayes net no terceiro experimento.

No experimento 1, nas primeiras semanas as taxas de acertos gerais ficaram entre 49,14% com o Simple Logistic e 54,30% com as redes neurais, enquanto que a acurácia de verdadeiros negativos ficou entre 47,75% e 55,29% com J48 e redes neurais respectivamente. Entretanto com o passar das semanas o algoritmo Bayes Net mantém um crescimento regular em ambas taxas de acerto com um pico de 75,96% na décima semana, tendência essa verificada até a semana 19. A partir deste momento o ramdon forest obteve os melhores resultados chegando a uma taxa de 97,88% de acerto no alunos classificados com o status de evadidos (verdadeiros negativos) e se mantendo praticamente assim até o final do curso.

No experimento 2, nas primeiras semanas de curso tivemos os melhores resultados com o algoritmo random forest variando entre 47.91% na primeira semana até 61,40% na semana 20. Com o passar das semana e o acréscimo dos dados as médias de acerto vão crescendo até o momento que atingem 96,69% na semana 76. A acurácia geral neste experimento foi muito próxima a de verdadeiros negativos exposta na figura 3, variando algo em torno de 1 a 2 pontos percentuais a cima da taxa de VN.

No terceiro experimento, que consistiu no treinamento com um curso e a aplicação dos modelos semanais nos dados de outro curso. Neste teste, foram gerados 103 modelos diferentes treinados com os dados do curso 1 e aplicados diretamente nos dados do curso 2. Isso tem como objetivo testar a aplicabilidade de modelos de predição de fácil generalização em outros cursos. Como podemos ver no gráfico, estes modelos necessitam de uma maior quantidade de dados para obterem resultados mais satisfatórios, entretanto, a partir da semana 11 esses resultados já atingem 63,12% de acurácia nos verdadeiros negativos e na semana 26 já obtém 91,26% de acerto.

No terceiro experimento foram alcançados os maiores percentuais de acerto na taxa de verdadeiros negativos 98,67%, resultado esse obtido utilizando o algoritmo J48 a partir da semana 51 do curso.

4. CONCLUSÕES

Como exposto neste artigo, diversos trabalhos na área de EDM buscam gerar modelos de predição da situação de alunos, entretanto, estes trabalhos são de difícil aplicação em outros casos se não os aplicados no estudos de origem. Desta forma, resolvemos criar modelos de fácil generalização e que pudessem ser aplicados em outros cursos e instituições alem da que foi utilizada como estudo de caso.

Como os resultados demonstram, a geração de modelos de predição para estudantes em risco em cursos Técnicos a Distância é possível e pode obter boas taxas de acurácia. Apesar dos resultados nas primeiras semanas de cursos serem baixos, temos de analisar que os cursos que foram testados são de longa duração e a evasão ocorre em todo o seu andamento. Assim, resultados obtidos

como o de 98,67% na semana 51 do terceiro experimento significam que na metade de um curso os modelos podem apontar quase a totalidade de alunos que irão se evadir. E até mesmo resultados como o de 91,26% na semana 26 que é antes do fim do primeiro quarto do curso demonstram a aplicabilidade dos modelos.

Vale ressaltar que este projeto ainda esta em desenvolvimento e a entre os trabalhos futuros podemos citar a expansão da base de dados com o acréscimo de outros cursos e também o teste dos modelos de predição em outros conjuntos de dados. Ainda que para facilitar o trabalho de pré-processamento e a geração dos modelos foi criado um sistema em Java que auxilia o processo e automatiza algumas das etapas, este ainda esta em desenvolvimento e aperfeiçoamento e fica como um trabalho futuro a sua finalização e total integração com a biblioteca do Weka e quem sabe diretamente com o banco de dados do ambiente virtual

5. REFERÊNCIAS BIBLIOGRÁFICAS

DELANO, R. and CORREA, S. “Redes na educação a distância: Uma análise estrutural do sistema uab em minas gerais,” **Revista PRETEXTO**., 2013.

SEGUNDO, F. R. and RAMOS, D. K., “Soluções baseadas no uso de software livre: alternativas de suporte tecnologico à educação presencial e a distancia,” Anais do 12 Congresso Internacional de Educação à Distancia, vol. 12, pp. 18–22, 2005.

BARROSO, M. F. and FALCAO E. B., “Evasão universitária: O caso do instituto de física da ufrj,” Encontro Nacional de Pesquisa em Ensino de Física, vol. 9, pp. 1–14, 2004.

MANHAES, L. M. B., CRUZ, S. d, COSTA, R. J. M., ZAVALETA J., and ZIMBRAO G., “Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados,” Anais do XXII SBIE-XVII WIE, Aracaju, 2011.

BAKER, R. S. J. D. and YACEF, K. “The State of Educational Data Mining in 2009 : A Review and Future Visions,” Journal of Educational Data Mining, vol. 1, no. 1, pp. 3–16, 2009.

BAKER, R., ISOTANI, S. and CARVALHO, A. Carvalho, “Mineração de Dados Educacionais: Oportunidades para o Brasil,” Revista Brasileira de Informatica na Educação , no. 02, p. 03, 2011.

QUEIROGA, E., CECHINEL, C., and ARAUJO, R., “Um Estudo do Uso de Contagem de Interações Semanais para Predição Precoce de Evasão em Educação a Distância,” Anais dos Workshops do IV Congresso Brasileiro de Informatica na Educação , p. 1074, 2015.