Predição de sucesso de estudantes cotistas utilizando algoritmos de classificação

Fellipe Henrique Mahon[1], Anderson Rufino dos Santos Silva[1], Daniel Miranda de Brito[1], Thaís Gaudencio do Rêgo[1]

[1]Centro de Informática – Universidade Federal da Paraíba (UFPB)
CEP 58051-900 – João Pessoa – PB – Brasil

{fellipe.mahon, andersonrufino2007, britmb, gaudenciothais}@gmail.com

Abstract. *In recent years, several educational institutions have adopted the quotas policy for admission. One problem is the performance difference between the quota and non-quota students. This paper aims to provide a method for identifying quota students with the possibility of having lower academic performance during the course, so that measures are taken to reduce the performance difference between the two groups. From tests using classifiers in Weka tool, with accuracy rates of up to 88%, it was found that the method may be a viable approach for identifying those groups of students.*

Resumo. *Nos últimos anos, diversas instituições de ensino têm adotado a política de cotas para ingresso. Um problema é a diferença de desempenho entre os cotistas e não cotistas. Este trabalho tem o intuito de fornecer um método para a identificação de estudantes cotistas com possibilidade de ter menor rendimento acadêmico durante o curso, de forma que medidas sejam tomadas para que as diferenças surgidas entre os dois grupos sejam reduzidas. A partir de testes com classificadores na ferramenta Weka, com taxas de acerto de até 88%, constatou-se que o método pode ser uma abordagem viável na identificação destes grupos de estudantes.*

1. Introdução

A adoção de políticas de ações afirmativas, principalmente as cotas, vem sendo aplicadas em diversas universidades públicas brasileiras como característica da política pública de inclusão social do governo federal, porém, ainda são constantes os debates e discussões na sociedade e no âmbito acadêmico a respeito de sua necessidade [Dallabona e Schiefler Filho 2011]. A viabilidade da adoção da política de cotas é objeto de estudo de várias pesquisas na área de educação, destacando-se aquelas que se concentram no rendimento e nas taxas de evasão dos alunos cotistas, conforme discutido a seguir.

Em um trabalho realizado com alunos da UnB, verificou-se que os alunos cotistas obtiveram rendimento substancialmente inferior ao dos alunos do sistema universal, apesar destes evadirem em proporções menores [Cardoso 2008]. Dallabona e Schiefler Filho (2011) encontraram diferenças de rendimento entre os alunos cotistas e não cotistas nos cursos da UTFPR, tendo os alunos cotistas rendimento inferior nas Engenharias e Bacharelados e pequena vantagem nos cursos de Licenciatura da mesma instituição. Em um estudo recente, englobando alunos de 15 cursos da Universidade

DOI: 10.5753/cbie.sbie.2015.997

Federal do Espírito Santo, constatou-se que o rendimento dos alunos cotistas é inferior em 9 dos 15 cursos avaliados [Pinheiro 2014]. Em outro trabalho [Peixoto et al. 2013] verificou-se que os alunos cotistas têm desempenho inferior em cursos de alta demanda social nas áreas de exatas e saúde. A situação muda quando se consideram os cursos de baixa demanda social, principalmente na área de artes e humanas. Mendes Junior (2013) concluiu em um estudo realizado com alunos da UERJ, que os cotistas possuem pior rendimento em relação aos não cotistas e que esta diferença é maior conforme a dificuldade relativa do curso, apesar disto, os cotistas evadem em taxas menores do que os não cotistas. É interessante, portanto, o desenvolvimento de políticas que permitam que as diferenças de desempenho entre os cotistas e não cotistas sejam reduzidas ou eliminadas, enquanto este tipo de política se fizer necessária, uma vez que a desigualdade social leva a esta demanda.

Técnicas de Mineração de Dados (do inglês, *Data Mining* - DM) têm sido utilizadas com sucesso na predição de desempenho de estudantes. A DM configura-se como uma abordagem para extração de conhecimento escuso de informações contidas em base de dados [Witten e Frank 2011] e faz parte de um processo mais amplo definido como Descoberta de Conhecimento em Bancos de Dados (do inglês, *Knowledge Discovery in Databases* - KDD). A aplicação das técnicas de DM em ambientes educacionais deu origem à área conhecida como Mineração de Dados Educacionais (do inglês, *Educational Data Mining* - EDM), que busca desenvolver metodologias para analisar conjuntos de dados oriundos de ambientes educacionais, simplificando a compreensão a respeito dos métodos de aprendizagem, além de outros fatores que influenciam diretamente a carreira acadêmica dos alunos [Baker et al. 2011].

O objetivo deste trabalho é fornecer um método para a identificação de alunos cotistas com risco de baixo rendimento durante o curso. Para investigar a eficiência do método utilizam-se dados de alunos de graduação cotistas da Universidade Federal da Paraíba (UFPB). O restante do artigo está organizado como segue: na seção 2, apresentam-se os Trabalhos Relacionados; na seção 3, discute-se a metodologia empregada; na seção 4, exibem-se os resultados e discussão; na seção 5, as considerações finais e trabalhos futuros são expostos.

2. Trabalhos Relacionados

Apesar da escassez de estudos na área de EDM, por se tratar de uma área ainda recente [Baker et al. 2011], a quantidade de trabalhos aplicados na previsão de desempenho dos estudantes cresce nos últimos anos, tanto na modalidade a distância (EaD), quanto no ensino presencial. Manhães et al. (2011) utilizou dados históricos de alunos concluintes e não concluintes de graduação do curso de Engenharia Civil da UFRJ para identificar os alunos do curso com risco de evasão. Foram utilizados como atributos de entrada as notas das disciplinas mais cursadas do primeiro período do curso no processo de aprendizagem.

Brito et al. (2014) propuseram um método para a predição do desempenho de estudantes do primeiro semestre do curso de Ciência da Computação da UFPB, baseado nas notas de ingresso no vestibular da instituição. Os autores acreditam que a identificação precoce dos estudantes pode ajudar educadores no combate a evasão, já que esta está muitas vezes relacionada a dificuldades nas disciplinas iniciais do curso.

Santos et al. (2012) utilizaram um método para a predição do desempenho de estudantes na avaliação somativa de uma disciplina com alta taxa de reprovação no curso de Engenharia da Computação da UNIPAMPA. Utilizaram-se como atributos de entrada para o classificador a presença do aluno em aulas, notas das avaliações formativas aplicadas no decorrer da disciplina em um ambiente Moodle. Gottardo et al. (2011) definiu um método para a previsão do desempenho final de alunos em uma disciplina em um curso EaD, baseado em dados de acesso ao ambiente de aprendizagem *Moodle*, como participação em fóruns, frequência de acesso ao sistema, entre outros. É notável a contribuição da área de EDM no aprimoramento dos sistemas de ensino, contudo verifica-se a ausência de trabalhos na área com foco no estudante cotista. Propõe-se neste trabalho, um método com foco neste tipo de estudante, de forma a enriquecer as discussões acerca do tema.

3. Metodologia

Os dados utilizados neste trabalho foram fornecidos pela Superintendência de Tecnologia da Informação (STI) da Universidade Federal da Paraíba (UFPB) e contém registros anônimos referentes aos seus alunos cotistas de graduação de todos os cursos presenciais, matriculados desde o segundo até o último período. Os alunos foram categorizados em dois grupos de acordo com o seu Coeficiente de Rendimento Escolar (CRE). Alunos com CRE superior a 7 foram considerados como sendo do grupo "Sucesso", enquanto alunos com CRE menor do que 7 foram considerados do grupo "Insucesso". Os atributos de entrada utilizados para a predição de sucesso dos alunos foram: "Forma de Ingresso na Instituição", "Curso", "Tipo de Cota" e "Nota de ingresso". Os três primeiros atributos são do tipo discretos e contém, respectivamente, 8, 116 e 15 valores possíveis e o atributo "Nota de Ingresso" é do tipo contínuo e pode assumir valores no intervalo 0-1000.

Para o estudo da viabilidade das predições foi utilizada a ferramenta Weka, escolhida por possuir diversos algoritmos de aprendizado de máquina implementados, permitindo que o usuário facilmente escolha o mais adequado ao problema. Com intuito de se obter o melhor resultado na classificação, foi realizada uma etapa de pré-processamento nos dados, eliminando a presença de instâncias com a ausência de algum dos atributos. De um total de 12971 instâncias do conjunto original, restaram 10130, sendo 5661 da classe "Sucesso" e 4469 da classe "Fracasso".

4. Resultados e Discussão

Para a avaliação do método proposto foram utilizados cinco algoritmos de aprendizagem de máquina pertencentes a classes distintas. Os parâmetros dos algoritmos não foram alterados e a divisão entre conjunto de treinamento e testes foi feita a partir de validação cruzada com 10 grupos, padrão da ferramenta.

Como medidas de avaliação dos algoritmos foram utilizadas a acurácia total, que representa a quantidade de instâncias classificadas corretamente e as taxas de verdadeiros e falsos positivos (Instâncias da classe "Fracasso" classificadas corretamente e incorretamente, respectivamente) e verdadeiros e falsos negativos (Instâncias da classe "Sucesso" classificadas corretamente e incorretamente, respectivamente). A Tabela 1 exibe os resultados dos classificadores testados.

Tabela 1. Avaliação dos classificadores

Métrica/ Classificador	J48	NaiveBayes	SMO	IBk	Multilayer Perceptron
Acurácia Média	81,09%	72,45%	71,06%	88,73%	67,07%
Verdadeiro Positivo	76,2%	65,7%	65,2%	87,3%	58,6%
Falso Negativo	23,8%	34,3%	34,8%	12,7%	41,4%
Verdadeiro Negativo	85%	77,8%	75,7%	89,9%	73,8%
Falso Positivo	15%	22,2%	24,3%	10,1%	26,2%

Na Tabela 1, verifica-se que as taxas de acerto (acurácia média) variam de 67,07% a 88,73%. É também importante observar as taxas de verdadeiros positivos que correspondem aos estudantes com provável baixo rendimento no curso, pois a identificação correta destes alunos é fundamental para a aplicação de medidas que visem a melhoria de rendimento deste grupo. O melhor algoritmo (IBk) conseguiu classificar corretamente os estudantes cotistas com risco de ter baixo rendimento (Verdadeiros Positivos) com taxa de 87,3% e os estudante cotistas com provável bom rendimento (Verdadeiros Negativos) com taxa de 89,9%, o que atesta o bom desempenho do algoritmo na tarefa de classificação. Avaliando ainda a importância de cada atributo para a tarefa de classificação, na ferramenta Weka, verificou-se que o "Tipo de Cota" é o menos relevante, o que revela que esse atributo não tem impacto no rendimento do aluno no curso, sendo os atributos "Curso" e "Nota de Ingresso" determinantes na classificação e, portanto, no desempenho do estudante. Os resultados obtidos confirmam a possibilidade de previsão do desempenho de estudantes em diversas estágios do curso, conforme discutido nos trabalhos relacionados. O fato de nenhum deles tratar de estudantes cotistas inviabiliza uma comparação detalhada dos resultados do trabalho.

5. Considerações Finais e Trabalhos Futuros

Neste trabalho, apresentou-se um método para a identificação de estudantes cotistas com possibilidade de ter baixo rendimento acadêmico durante o curso. Utilizaram-se dados de alunos cotistas de todos os cursos de graduação presenciais da UFPB para demonstrar a possibilidade de identificação destes estudantes. Apesar do corrente estudo ainda estar em fase inicial, demonstra um grande potencial na previsão do sucesso dos alunos cotistas da instituição, a partir de um conjunto pequeno de atributos, onde o tipo de cota é o atributo que menos influencia no sucesso ou fracasso do estudante, o que sugere que a situação econômica, raça e tipo de ensino (público ou privado) cursado não tem tanto impacto no desempenho final do estudante. A eficácia do método na identificação da situação dos estudantes foi atestada através da aplicação de classificadores na ferramenta Weka.

Espera-se que os resultados obtidos neste trabalho possam ser utilizados pela instituição na identificação precoce de estudantes cotistas com possibilidade de ter menor desempenho na graduação. A identificação destes estudantes pode ajudar no

planejamento de estratégias para tentar melhorar o rendimento acadêmico destes alunos no decorrer do curso, de forma a minimizar as diferenças de rendimento em relação aos não cotistas. Os bons resultados obtidos neste trabalho motivam pesquisas futuras, como a continuidade do atual estudo com o intuito de se obter resultados mais detalhados acerca do impacto da cota no desempenho dos estudantes cotistas, considerando também o acréscimo dos estudantes não cotistas no estudo.

Referências

Baker, R. S. J., Isotani, S. e Carvalho, A. M. J. B. (2011). Mineração de dados educacionais: Oportunidades para o Brasil. Revista Brasileira de Informática na Educação, v.19, n.2, p. 3-13.

Brito, D. M., de Almeida Júnior, I. A., Queiroga, E. V. e do Rêgo, T. G. (2014). Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina. In *Anais do Simpósio Brasileiro de Informática na Educação*, v. 25, n.1, p. 882 – 890.

Cardoso, C. B. (2008). Efeitos da política de cotas na Universidade de Brasília: uma análise do rendimento e da evasão. Dissertação de Mestrado.

Dallabona, C. A. e Schiefler Filho, M. F. D. O (2011). Desempenho Acadêmico de estudantes oriundos de escolas públicas: cursos de graduação do campus Curitiba da UTFPR. In: *XXXIX Congresso Brasileiro de Educação em Engenharia*, p. 2040-2051.

Gottardo, E., Kaestner, C. e Noronha, R. V. (2012). Previsão de Desempenho de Estudantes em Cursos EAD Utilizando Mineração de Dados: uma Estratégia Baseada em Séries Temporais. In *Anais do Simpósio Brasileiro de Informática na Educação*, v. 23, n.1.

Manhães, L. M. B., Cruz, S. D., Costa, R. J. M., Zavaleta, J., e Zimbrão, G. (2011). Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. In: *Anais do XXII SBIE-XVII WIE, Aracaju*, v. 22, n. 1, p. 150-159.

Mendes Junior, A. A. F. (2013). Uma análise da progressão dos alunos cotistas sob a primeira ação afirmativa brasileira no ensino superior: o caso da Universidade do Estado do Rio de Janeiro (UERJ). *Texto Para Discussão*, n. 71, p. 1-23.

Peixoto, A. D. L. A., Ribeiro, E. M. B. D. A., Bastos, A. V. B. e Ramalho, M. C. K. (2013). Cotas e desempenho acadêmico na UFBA: Um estudo a partir dos coeficientes de rendimento. In: *XIII Coloquio de Gestión Universitaria en Américas*.

Pinheiro, J. S. S. P. (2014). Desempenho acadêmico e sistema de cotas: um estudo sobre o rendimento dos alunos cotistas e não cotistas da Universidade Federal do Espirito Santo. Dissertação de Mestrado

Santos, H., Camargo, F. e Camargo, S. (2012). Minerando Dados de Ambientes Virtuais de Aprendizagem para Predição de Desempenho de Estudantes. *Conferencias LACLO*, v. 3, n. 1.

Witten, I. H. e Frank, E. (2011). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 3ª edição.