Universidade Federal do Ceará

Centro de Ciências

Departamento de Computação

Programa de Pós-Graduação em Ciência da Computação

Abelardo Vieira Mota

Aplicação de aprendizado de máquina ao problema de evasão de discentes da UFC

Fortaleza

2015

Abelardo Vieira Mota

Aplicação de aprendizado de máquina ao problema de evasão de discentes da UFC

João Paulo Pordeus Gomes

Fortaleza

2015

Nem terminei ainda rs

"Prefiro ser
Essa metamorfose ambulante"
(Raul Seixas)

Resumo

The ideal abstract will be brief, limited to one paragraph and no more than six ou seven sentences, to let readers scan it quickly for an overview of the paper's content.

Palavras-chaves: Aprendizado de máquina. Evasão.

Abstract

I don't speak english.

Key-words: Machine Learning. Drop-out.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

(BAKER; ISOTANI; CARVALHO, 2011) Em estudo realizado no departamento de engenharia elétrica da Eindhoven University of Technology(DEKKER; PECHENIZKIY; VLEESHOUWERS, 2009), é relatado que em dezembro os discentes desse departamente recebem um aviso informando se são ou não aconselhados a continuarem no curso. Esse aviso é baseado na performance do discente no curso e em informações obtidas de professores do primeiro semestre e de discentes monitores. É relatado que o aviso parece ter bastante acurácia: geralmente discentes aconselhados a continuarem têm sucesso no próximo ano do curso, enquanto aqueles desaconselhados geralmente não continuam no curso. Nesse estudo foram utilizados diversos algoritmos de aprendizado de máquina com o objetivo de tentar detectar que um estudante irá abandonar seu curso. Foram utilizadas informações de discente referentes tanto ao período anterior ao seu ingresso na universidade, quanto ao posterior.

Uma das estratégias adotadas para diminuir as taxas de evasão é a identificação precoce de discentes com grande tendência para abandonarem seus cursos e a execução de ações que minimizem tal tendência. A identificação pode ser conduzida por observação do comportamento e resultados dos discentes, de forma subjetiva, pelos docentes e coordenadores de cursos, por exemplo. Dois problemas decorrem dessa forma de identificação: sendo conduzida por pessoas, essa forma de identificação é limitada pelo conjunto de observações as quais o observador tem acesso; sendo subjetiva, seus resultados podem sofrer resistência para serem aceitos. A utilização de técnicas de aprendizado de máquina como forma de identificação pode contornar esses problemas, por, primeiro, fazer uso de dados registrados por sistemas de informação, provavelmente contendo informações mais amplas que as que uma pessoa pode observar; segundo, por fazer maior uso de dados registrados, sendo aceita mais facilmente como identificação objetiva.

O Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais(REUNI), instituído pelo DECRETO Nº 6.096, DE 24 DE ABRIL DE 2007 , possui como uma de suas diretrizes:

> I - redução das taxas de evasão, ocupação de vagas ociosas e aumento de vagas de ingresso, especialmente no período noturno;

Na Universidade Federal do Ceará(UFC), de acordo com o Anuário estatístico de 2014, ano base 2013, o indicador "Taxa de sucesso na graduação", definido como a proporção entre número de discentes diplomados e número de discentes ingressantes da graduação, esteve em 2013 com o menor valor desde 2008(Figura 1). Já o indicador "Taxa

de sucesso da graduação por curso", em 2013, possuiu valor mínimo igual a 6.8%, referente ao curso Ciências Sociais, habilitação em licenciatura(Tabela 1).

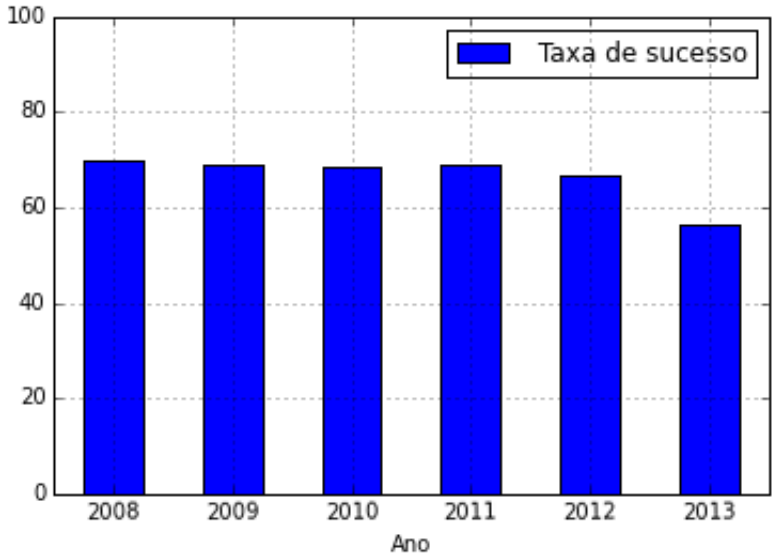

Figura 1 – Taxa de Sucesso na Graduação - UFC

Tabela 1 Taxa de sucesso da graduação por curso na UFC em 2013 - 10 piores resultados, em ordem crescente

Curso	Período	Taxa de Sucesso
Ciências Sociais - Licenciatura	Noturno	6.8%
Redes de Computadores - Quixadá	Noturno	13.3%
Geografia - Bacharelado	Diurno	15.3%
Letras - Português-Alemão	Diurno	17.6%
Engenharia Metalúrgica	Diurno	18.3%
Ciências Econômicas - Sobral	Noturno	20.9%
Sistemas de Informação - Quixadá	Diurno	22.0%
Filosofia - Bacharelado	Noturno	24.3%
Matemática - Bacharelado	Diurno	24.4%
Engenharia Elétrica - Sobral	Diurno	25.0%

Em DIRETRIZES GERAIS DO PROGRAMA DE APOIO A PLANOS DE REESTRUTURAÇÃO E EXPANSÃO DAS UNIVERSIDADES FEDERAIS REUNI p.4 é ressaltado que o indicador "Taxa de conclusão dos cursos de graduação", cuja definição é igual à do "Taxa de Sucesso na graduação":

> (...) não expressa diretamente as taxas de sucesso observadas nos cursos da universidade, ainda que haja uma relação estreita com fenômenos de retenção e evasão.

Esse indicador não é sensível, por exemplo, à ocorrência de uma greve, fenômeno que causa atraso na formação dos discentes, podendo diminuir relevantemente a quantidade de diplomados em determinado ano. Funciona bem considerando o modelo em que, necessariamente, após a diplomação de uma turma de discentes, turma de igual tamanho irá ingressar. A realidade mostra que o processo de diplomação, iniciado com o ingresso do discente e finalizado com a emissão e recebimento de seu diploma, é mais complexo. Se adicionarmos o indivíduo que assume o papel de discente, a análise das causas da evasão torna-se mais complexa, trazendo à tona dimensões como a social e a econômica, além da vida pré universidade do indivíduo. Em (VILLWOCK; APPIO; ANDRETA, 2015), por exemplo, informações sobre o trabalho do discente e se ele é casado foram os fatores mais relevantes na classificação dos discentes analisados com relação à evasão do curso.

O presente trabalho objetiva avaliar a aplicabilidade de técnicas de aprendizado de máquina ao problema de evasão de discentes na UFC a partir dos dados que seus sistemas de informação gerenciam. A UFC possui uma base de dados de informações sobre seus discentes gerada e mantida pelo sistema SIGAA(Sistema Integrado de Gestão de Atividades Acadêmicas). Para tanto é necessário que seja feito uma análise sobre a estrutura e qualidade dos dados disponíveis.

2 Aprendizado de máquina

Aprendizado de máquina é uma subárea de Inteligência Artificial que agrupa conhecimentos sobre algoritmos e técnicas que permitam que um programa melhore sua performance a partir de dados. Mais formalmente, (MITCHELL, 1997, p.2, tradução nossa) define:

> Um programa aprende a partir de uma experiência E, com relação a uma classe de tarefas T e a uma medida de performance P, se sua performance em tarefas da classe T, medida por P, melhora com a experiência E.

Seja, por exemplo, o problema de identificação de autoria de textos. Uma das abordagens utilizada para resolvê-lo é verificar a similaridade entre o texto em análise e um conjunto de textos cujos autores já sejam conhecidos, denominado conjunto de treino, sendo reportado como o autor aquele cujos textos contidos no conjunto de treino sejam mais similares ao texto em análise. Nesse exemplo, o elemento experiência é representado por um conjunto de textos rotulados com seus respectivos autores; o elemento tarefa é a identificação do autor de um texto; o elemento medida de performance é a proporção de textos cujos autores são corretamente identificados.

(MITCHELL, 1997), para detalhar a modelagem de um programa com uma abordagem de aprendizado de máquina, apresenta uma sequência de passos para desenvolver um programa que aprenda a jogar xadrez, que aqui adaptamos para o problema de evasão de discentes.

O primeiro passo é a escolha da medida de performance. Um exemplo é a quantidade de discentes identificados que efetivamente abandonarem seus cursos. Uma questão surge: se o programa a ser desenvolvido for ser utilizado para identificar discentes que potencialmente abandonarão seus cursos, para os quais serão realizadas atividades para diminuição desse potencial, como contar precisamente quantos deles foram corretamente identificados? Como verificar que um discente que tinha a pretenção de abandonar o curso no momento da identificação pelo programa foi corretamente identificado se, após ações realizadas com a finalidade de diminuir seu potencial de evasão, ele concluiu o curso? A identificação pelo programa, seguido das ações, torna-se mais uma variável concorrente para o resultado do discente no curso. Outra questão relevante é o custo dos erros na identificação: considerando que identificar corretamente um discente com potencial de evasão implica em diminuição desse potencial, e que identificar incorretamente um discente sem potencial de evasão não implica em aumento desse potencial, concluímos

que a performance do programa deve levar em consideração a diferença entre os custos de seus erros.

De acordo com (MITCHELL, 2006), o conhecimento sobre aprendizado de máquina pode ser aplicado a diversas áreas, como, por exemplo, a reconhecimento de voz; a visão computacional, sendo utilizado no desenvolvimento de sistemas de reconhecimento facial; a controle de robôs.

(SCULLEY et al., 2014) (DOMINGOS, 2012)

3 Evasão de discentes

Em (DEKKER; PECHENIZKIY; VLEESHOUWERS, 2009) são aplicados algoritmos de aprendizado de máquina a dados de discentes do Electrical Engineering department, Eindhoven University of Technology, considerando o período de 2000 a 2009, com o objetivo de identificar discentes em grupos de risco de evasão. É relatado que esse departamento já avaliava os discentes com relação ao risco de evasão, mas de forma subjetiva. O estudo ressalta o maior custo da ocorrência de falsos negativos que de falsos positivos na identificação de discentes com risco de evasão. Ocorre que, argumenta-se, há prejuízo maior em não oferecer apoio a um discente com risco de evasão do que oferecer, desnecessariamente, apoio a um discente sem tal risco. O estudo faz uso então de uma matriz de custo, com o classificador CostSensitiveClassifier, do Weka, obtendo melhores resultados, com relação a ocorrência de falsos negativos, mas com perdas de acurácia.

Em (MANHÃES; CRUZ; ZIMBRÃO,) são aplicados algoritmos de aprendizado de máquina a dados de discentes de seis cursos da Universidade Federal do Rio de Janeiro, com o objetivo de identificar discentes que não terão pelo menos uma aprovação no segundo semestre de seus cursos. Os cursos considerados foram: Direito, Farmácia, Física, Engenharia Civil, Engenharia Mecânica, Engenharia de Produção. É indicado que esses cursos foram escolhidos por pertencerem a departamentos distintos, com perfis de discentes distintos. É observado também que tais cursos diferem com relação à quantidade de discentes ingressantes, à taxa de evasão registrada e à efetividade de certas práticas de ensino. Para cada curso são desenvolvidas uma base de dados de treinamento e uma base de dados de teste, composta pelos dados de seus discentes de primeiro semestre. Para a base de dados de treinamento foram utilizados os dados dos anos pares(de 1994.1 a 2008.1), já para a de teste foram utilizados os dados dos anos ímpares(de 1995.1 a 2009.1). Foram utilizados os algoritmos Naive Bayes, Multilayer Perceptron, Support Vector Machine e Decision Tree.

Definições de evasão:

- após prazo T conseguiu performance P

Tabela 2 Definição de evasão

Artigo	Período(em semestres)	Medida de performance
(DEKKER; PECHENIZKIY; VLEESHOUWERS, 2009)	6	Se conseguiu a propedeusa

Tabela 3 Tamanho do dataset

Artigo	Quantidade de instâncias	Período
(DEKKER; PECHENIZKIY; VLEESHOUWERS, 2009)	648	2000-2009
	495	2000-2009
	516	2000-2009
	516	2000-2009

Características da base de dados:

4 Pontos de partida

Como pontos de partida, consideraremos:

- Analisar o problema da predição de aluno com alta probabilidade de evasão do curso a partir de dados referentes a sua vida pré universidade e dados de seu desempenho no primeiro semestre do curso. Esse problema é relevante àqueles interessados na diminuição de taxas de evasão de cursos a partir de atividades voltada a discentes em grupo de risco de evasão.
- Analisar o problema da predição do desempenho de um discente em uma disciplina a partir de dados sobre seu histórico como discente. Esse problema é relevante àqueles interessados na melhoria do desempenho de discentes em uma determinada disciplina a partir de atividades a serem iniciadas antes de o discente efetivamente cursá-la.
- Analisar o problema da predição da taxa de evasão de um curso a partir de seus dados, como, por exemplo, de sua estrutura curricular. Esse problema é relevante ao processo de desenho ou redesenho de um curso, sua solução podendo ser utilizado como guia para decisões acerca do curso.

Referências

BAKER, R. S. J. de; ISOTANI, S.; CARVALHO, A. M. J. B. de. Mineraçao de dados educacionais: Oportunidades para o brasil. 2011.

DEKKER, G. W.; PECHENIZKIY, M.; VLEESHOUWERS, J. M. Predicting students drop out: A case study. *International Working Group on Educational Data Mining*, ERIC, 2009.

DOMINGOS, P. A few useful things to know about machine learning. *Communications of the ACM*, ACM, v. 55, n. 10, p. 78–87, 2012.

MANHÃES, L. M. B.; CRUZ, S. M. S. da; ZIMBRÃO, G. Evaluating performance and dropouts of undergraduates using educational data mining.

MITCHELL, T. M. *Machine Learning*. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072.

MITCHELL, T. M. *The discipline of machine learning*. [S.l.]: Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006.

SCULLEY, D. et al. Machine learning: The high interest credit card of technical debt. In: *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*. [S.l.: s.n.], 2014.

VILLWOCK, R.; APPIO, A.; ANDRETA, A. A. Educational data mining with focus on dropout rates. *International Journal of Computer Science and Network Security (IJCSNS)*, International Journal of Computer Science and Network Security, v. 15, n. 3, p. 17, 2015.