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

O problema de evasão de discentes consiste no abandono, pelo discente, de um processo de estudos antes de sua conclusão. Essa definição pode ser detalhada especificando o escopo do processo de estudos: a evasão pode ser de um curso, de uma instituição de ensino, de cursos de uma determinada área, do sistema de ensino etc. O abandono do discente sem a conclusão do processo de estudos, considerando a finalidade desse processo, representa desperdício de recursos e de tempo de todos os envolvidos: discente, docentes, instituição de ensino e sociedade.

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).

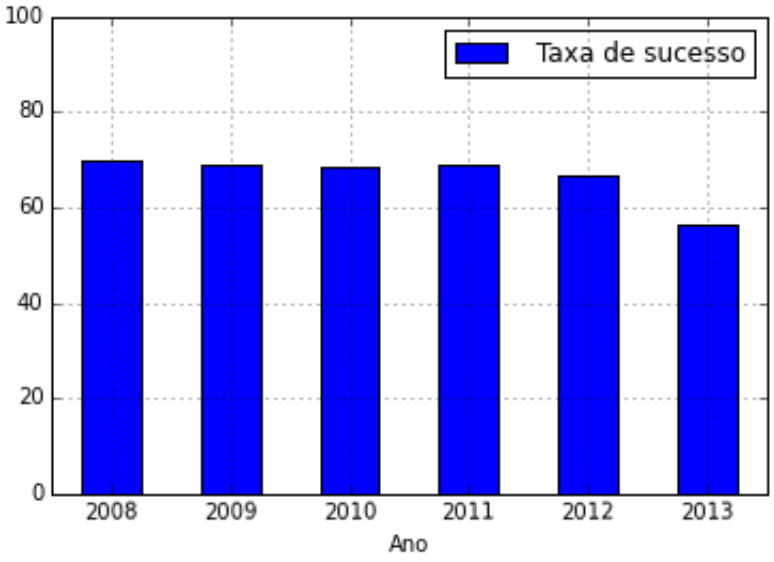

Figura 1 – Taxa de Sucesso na Graduação - UFC

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.

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. 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. 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. 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.

Na UFC o estudo (ANDRIOLA; ANDRIOLA; MOURA, 2006) foi desenvolvido e objetivou analisar o problema da evasão de discentes através da análise da opinião de docentes e coordenadores sobre esse problema. 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, a ser utilizado para disputar um campeonato mundial de xadrez. É escolhida como medida de performance a quantidade de vitórias do programa nesse campeonato.

O primeiro passo é a escolha da experiência a partir da qual o programa irá aprender, denominada experiência de treinamento. (MITCHELL, 1997) classifica os tipos de experiência a partir de três atributos: feedback da experiência, se direto ou indireto com relação a como o programa será utilizado; nível de controle que há sobre a experiência; e quão bem a experiência reflete a realidade. Ressalta que o tipo de experiência utilizada pelo programa pode ter impacto significativo no sucesso ou falha em seu aprendizado.

Com relação ao feeedback, o elemento experiência pode ser classificado em experiência de feedback direto e experiência de feedback indireto. Por exemplo, a tupla estado do tabuleiro e melhor movimento possível é classificada como experiência de feedback direto: o programa irá atuar realizando jogadas e esse tipo de experiência informa diretamente se uma jogada é boa ou não. Já a tupla sequência de jogadas de uma partida e seu resultado final é classificada como experiência de feedback indireto: o programa deverá inferir a qualidade de uma jogada a partir do resultado de uma partida da experiência em

que ela apareça e de seu resultado. À atividade de determinar o grau de influência que elementos de uma experiência de feedback indireto têm sobre o resultado é denominada credit assignment.

Com relação ao nível de controle, o elemento experiência pode ser classificado em experiência selecionada por especialista, experiência sugerida pelo programa e analisada por um especialista e experiência selecionada e analisada pelo programa. Por exemplo, a experiência será do tipo selecionada por especialista se houve um jogador experiente de xadrez que selecione estados de tabuleiro e indique que melhores jogadas poderão ser feitas; será do tipo sugerida pelo programa e analisada por um especialista se o próprio programa selecionar estados de tabuleiro para serem analisadas por um jogador experiente de xadrez; será do tipo selecionada e analisada pelo programa se o programa utilizar o resultado de partidas que disputar consigo mesmo.

Com relação a quão bem reflete a realidade, o elemento experiência pode ser classificado se sua distribuição representa a distribuição dos exemplos com os quais o programa efetivamente será utilizado. Por exemplo, a experiência não irá refletir a realidade caso esteja limitada ao conjunto de partidas de apenas um jogador: considerando que o programa será utilizado em um campeonato mundial, do qual participam jogadores diversos, com estilos de jogo diversos, é capaz de o programa, com essa experiência treinado, depare com estados de tabuleiro que não encontrou no treinamento. (MITCHELL, 1997) ressalta que muito da teoria de aprendizado de máquina depende da assunção de que a experiência utilizada no treinamento reflete a realidade.

O próximo passo é a escolha do tipo de conhecimento que deverá ser aprendido, representado por uma função denominada função alvo, e como ele será utilizado pelo programa. Considerando que o programa irá ser utilizado como um jogador de xadrez, um tipo de conhecimento que pode ser escolhido é uma função que tenha como domínio o conjunto de estados de tabuleiro e retorne a melhor jogada a partir de um dado estado de tabuleiro informado. O programa irá jogar realizando a jogada que essa função retornar, a partir do estado atual do tabuleiro. Esse tipo de conhecimento, apesar de atraente, é tão difícil de ser adquirido quão difícil for determinar quanto uma jogada influencia no resultado final de uma partida. Outro tipo de conhecimento é uma função que tenha como domínio o conjunto de estados de tabuleiro e retorne um número real, indicando quão bom o estado de tabuleiro informado é. O programa irá jogar verificando qual estado de tabuleiro maximiza o valor da função, considerando o conjunto de estados de tabuleiro que podem ser alcançados a partir do estado atual do tabuleiro e de todas jogadas válidas.

Após a escolha do tipo de conhecimento que deverá ser aprendido, é necessário definir como esse conhecimento será representado. A função que associa um estado de tabuleiro a um número real pode assumir diversas formas: pode ser uma matriz contendo uma célula com um número real para cada estado de tabuleiro possível; pode ser um

conjunto de regras que associe atributos do estado do tabuleiro a números reais; pode ser uma função polinomial de atributos do estado do tabuleiro em números reais etc. Para darmos continuidade ao detalhamento dos passos, escolhemos aqui uma representação de função simples: denominaremos por V a função que associa um estado de tabuleiro a um número real, calculada como combinação linear dos seguintes atributos do estado do tabuleiro:

- x_1: número de peças pretas no tabuleiro
- x_2: número de peças brancas no tabuleiro
- x_3: número de reis pretos no tabuleiro
- x_4: número de reis brancos no tabuleiro
- x_5: número de peças pretas ameaçadas por peças brancas no tabuleiro
- x_6: número de peças brancas ameaçadas por peças pretas no tabuleiro

A função V pode ser representada então por:
$V(t) = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 + w_6x_6$
Os coeficientes w_0 a w_6 são parâmetros do programa, a serem ajustados no aprendizado. Resumindo os passos até aqui realizados, temos:

- Tarefa: jogar xadrez
- Medida de performance: proporção de jogos do campeonato mundial de xadrez ganhos
- Experiência de treinamento: partidas disputadas pelo programa contra ele mesmo
- Função alvo: $V : EstadosDeTabuleiro \rightarrow \mathbb{R}$
- Representação da função alvo: $V(t) = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 + w_6x_6$

O próximo passo consiste na escolha de um algoritmo que, a partir de um conjunto de experiências, irá ajustar os parâmetros da representação da função a fim de aproximá-la da função alvo. Para tanto, é necessário um conjunto de dados de treinamento, composto por um par de estado de tabuleiro e o valor que a função deverá atribuir. Denotamos por $\langle b, V_{treino}(b)\rangle$ um dado para treinamento: b é uma tupla contendo os atributos de um estado de tabuleiro e $V_{treino}(b)$ o valor a ele atribuído. Por exemplo:

$\langle\langle x_1 = 3, x_2 = 0, x_3 = 1, x_4 = 0, x_5 = 0, x_6 = 0\rangle, +100\rangle$

Determinar os valores dos estados de tabuleiros utilizados no treinamento é tarefa fácil para os estados finais: pode-se definir dois valores, $max < min$, e atribuir aos estados finais de vitória o valor max e aos de derrota o valor min. Já a determinação dos valores dos tabuleiros intermediários não é tão simples. O fato de uma partida ter sido ganha não implica necessariamente que todos os estados de tabuleiro nela percorridos devem receber um valor alto. Para definir tais valores pode-se utilizar uma regra de estimação. Por exemplo:

$$V_{treino}(b) \leftarrow V(sucessor(b))$$

Essa regra atribui a um estado de tabuleiro de treinamento o valor que a função alvo estimada retorna para o estado de tabuleiro após a próxima jogada do oponente. Apesar de parecer estranho fazer uso de V, a função que se está estimando, para determinar os valores a serem utilizados para refiná-la, intuitivamente essa abordagem parece fazer sentido, por atribuir a um estado de tabuleiro um valor que é função de um estado que foi possível alcançar a partir dele.

Após a determinação dos valores dos estados de tabuleiros contidos na experiência de treinamento, o conjunto de dados de treinamento pode ser utilizado. Para estimar a função alvo a partir dos dados de treinamento, dada a representação da função escolhida, é necessário agora determinarmos seus pesos, w_0 a w_6. Para tanto, primeiro definimos como mensurar quão bem a função estimada se adequa aos dados de treinamento. Uma medida comum é o erro quadrático, assim definido:

$$E = \sum_{\langle b, V_{treino}(b)\rangle \in \ dados\ de\ treinamento} (V_{treino}(b) - V(b))^2$$

O problema de estimar a função alvo pode ser modelado então como o problema de encontrar os pesos w_0 a w_6 que minimizem o erro quadrático sobre os dados de treinamento. Um dos algoritmos que incrementalmente ajusta os pesos aos dados de treinamento, minimizando o erro quadrático é o Método dos Mínimos Quadrados. Esse algoritmo funciona ajustando, para cada dado de treinamento, os pesos da função na direção que minimiza o erro quadrático. Para tanto, atualiza iterativamente, para cada dado de treinamento, os pesos da função, incrementando com um valor proporcional a $(V_{treino}(b) - V(b))$: se o valor da função aplicado ao dado de treinamento, $V(b)$, é igual ao valor do dado de treinamento, $V_{treino}(b)$, o incremento será nulo; se o valor da função é maior que o do dado de treinamento, o incremento será negativo, fazendo com que, para o dado de treinamento em questão, o valor da função diminua; se o valor da função é menor que o do dado de treinamento, o incremento será positivo, fazendo com que, para o dado de treinamento em questão, o valor da função aumente.

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.

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

ANDRIOLA, W. B.; ANDRIOLA, C. G.; MOURA, C. P. Opiniões de docentes e de coordenadores acerca do fenômeno da evasão discente dos cursos de graduação da universidade federal do ceará (ufc). *Ensaio: aval. pol. públ. Educ*, SciELO Brasil, 2006.

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

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.

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.