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

Abelardo Vieira Mota

25 de maio de 2015

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.

Lista de Figuras

Sumário

1 Introdução

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

como cita isso?

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

Na UFC, de acordo com o último Anuário estatístico, de 2014, ano base 2013, o indicador "Taxa de Sucesso na Graduação", definido como a proporção entre número de diplomados e número de ingressantes da graduação, esteve em 2013 com o menor valor desde que passou a ser monitorado. Já o indicador "Taxa de sucesso da graduação por curso", em 2013, possuiu valor médio igual a 64% e valor mínimo igual a 6.8%(Tabela 1).

na ufc

quando sai o próximo?

qual a definição?

Uma das estratégias para diminuir as taxas de evasão é a identificação precoce de estudantes 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 alunos, de forma subjetiva, pelos docentes e coordenadores de cursos, por exemplo. Dois problemas decorrem dessa forma de identificação: sendo conduzida por pessoas, essa

como resolve

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%

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

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.

Em estudo realizado no departamento de engenharia elétrica da Eindhoven University of Technology[3], é relatado que em dezembro os estudantes desse departamente recebem um aviso informando se são ou não aconselhados a continuarem no curso. Esse aviso é baseado na performance do estudante no curso e em informações obtidas de professores do primeiro semestre e de estudantes monitores. É relatado que o aviso parece ter bastante acurácia: geralmente estudantes 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.

é o que

outros estudos

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)

dados na ufc

O presente trabalho objetiva avaliar a aplicabilidade de técnicas de mineração de dados sobre o problema de evasão de estudantes na UFC a partir dos dados que seus sistemas de informação gerenciam.

referencia

Para tanto é necessário que seja feito uma análise sobre a estrutura e qualidade dos dados disponíveis.

Também serão levantadas hipóteses sobre causas para o problema e será analisado se os dados as corroboram ou não.

hipótese sobre a estrutura do curriculo -> métricas com relação às turmas! por exemplo, distância de horário entre disciplinas do mesmo semestre -> no caso do aluno, para cada semestre calcular

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, [7] define que 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 autoria de textos, que consiste na identificação automatizada do autor de um texto. Uma das soluções desenvolvidas é a verificação da 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 um conjunto de textos rotulados com seus respectivos autores; o elemento tarefa é a identificação do autor de um dado texto; o elemento medida de performance a proporção de textos cujos autores são corretamente identificados.

[9] [4]

problemas com tarefa, medida de performance, experiencia menos comuns

3 Evasão de estudantes

Em [3] são aplicados algoritmos de aprendizado de máquina a dados de estudantes do Eletrical Engineering department, Eindhoven University of Technology , com o objetivo de identificar estudantes em grupos de risco de evasão. É relatado que esse departamento já avaliava os estudantes 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 estudantes com risco de evasão. Ocorre que, argumenta-se, há prejuízo maior em não oferecer apoio a um estudante com risco de evasão do que oferecer, desnecessariamente, apoio a um estudante 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.

[8] [10] [1] [11] [6] [5] [2]

bla bla bla de machine learning

definição informal de aprendizado de máquina

definição formal de aprendizado de máquina

métricas

algoritmos

métodos de teste

aplicação

machine learning

tabela com dados dos estudos analisados - ver fichamento.xlsx

Referências

[1] RSJD Baker et al. Data mining for education. *International encyclopedia of education*, 7:112–118, 2010.

[2] Ryan Shaun Joazeiro de Baker, Seiji Isotani, and Adriana Maria Joazeiro Baker de Carvalho. Mineraçao de dados educacionais: Oportunidades para o brasil. 2011.

[3] Gerben W Dekker, Mykola Pechenizkiy, and Jan M Vleeshouwers. Predicting students drop out: A case study. *International Working Group on Educational Data Mining*, 2009.

[4] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[5] Laci Mary Barbosa Manhães, SMS Cruz, Raimundo J Macário Costa, Jorge Zavaleta, and Geraldo Zimbrão. Identificaçao dos fatores que influenciam a evasao em cursos de graduaçao através de sistemas baseados em mineraçao de dados: Uma abordagem quantitativa. *Anais do VIII Simpósio Brasileiro de Sistemas de Informação, São Paulo*, 2012.

[6] Laci Mary Barbosa Manhães, Sérgio Manuel Serra da Cruz, and Geraldo Zimbrão. Evaluating performance and dropouts of undergraduates using educational data mining.

[7] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[8] Cristóbal Romero and Sebastián Ventura. Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6):601–618, 2010.

[9] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, and Michael Young. Machine learning: The high interest credit card of technical debt. In *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*, 2014.

[10] Brandon Sherrill, William Eberle, and Doug Talbert. Analysis of student data for retention using data mining techniques. 2011.

[11] Rosangela Villwock, Andressa Appio, and Aldioni Adaiani Andreta. Educational data mining with focus on dropout rates. *International Journal of Computer Science and Network Security (IJCSNS)*, 15(3):17, 2015.