Aplicação de mineração de dados ao problema de evasão de estudantes da UFC

Abelardo Vieira Mota

24 de maio de 2015

Resumo

resumo

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:

> 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"(TS) , 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 mínimo de 11.7% para o curso Letras Português-Alemão, considerando

necessidade, tendência

como cita isso?

DECRETO Nº 6.096, DE 24 DE ABRIL DE 2007

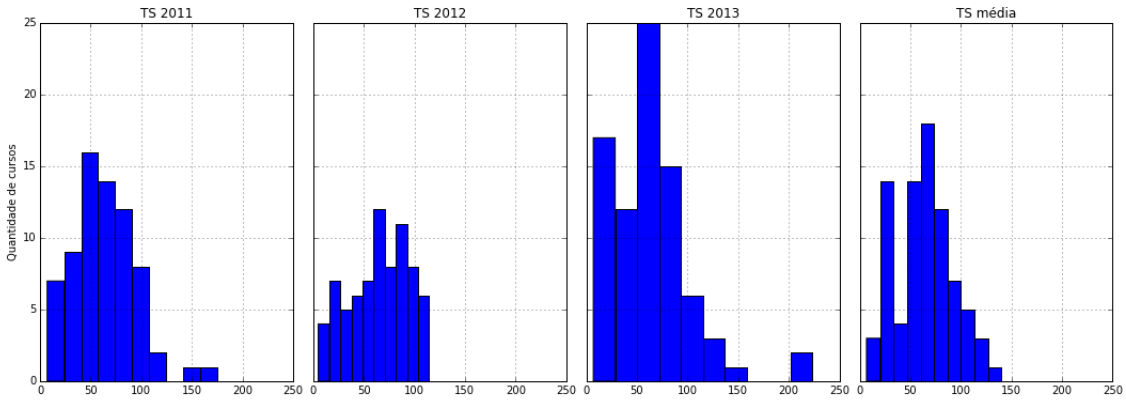
na ufc

quando sai o próximo?

figura

qual a definição?

Figura 1: Histogramas da medida taxa de sucesso na UFC



a média entre os anos 2013, 2012 e 2011, para os cursos que tiveram valores calculados nesses anos.

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 decorrer 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 mineração de dados 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 docente, por exemplo, pode observar; segundo, por fazer maior uso de dados registrados, sendo aceita mais facilmente como identificação objetiva. Dessa forma, podem colaborar com a solução do problema, seja auxiliando na identificação por pessoas, seja realizando a identificação automaticamente.

como resolve

dados do anuário mais detalhados

Os dados gerenciados pelos sistemas de informação da UFC podem conter informações que auxiliem o entendimento das causas desse problema e permitam que melhor sejam planejadas ações para solucioná-lo ou que estudantes com maior risco de evasão sejam identificados e recebam apoio da instituição. É nesse sentido, que a área de pesquisa denominada mineração de dados(data mining, em inglês), que estuda como extrair informações sobre dados, pode contribuir, fornecendo meios para a descoberta de informações relevantes a partir dos dados registrados pela UFC.

dados na ufc

falar sobre os estudos sobre drop-out+data mining

1.1 Objetivos

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.

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

1.2 Organização do trabalho

no fim

2 Mineração de dados

sobre data mining - qual o super artigo básico?

2.1 Identificação de padrões

clustering, visualização

2.2 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 o aprendizado de um programa por

como por a definition?

Definition: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Para a aplicação dos conhecimentos de aprendizado de máquina para solucionar um problema, esses três elementos, tarefa(task), medida de performance(performance measure) e experiência(experience) devem ser identificados.

Reconhecimento de dígito manualmente escrito

por que não aplicar outras técnicas?

O problema de reconhecimento de dígito manualmente escrito consiste na identificação automatizada do valor de um dígito manualmente escrito contido em uma imagem.

exemplos

Tarefa: dada uma imagem contendo um dígito manualmente escrito, identificar qual o valor do dígito nela contido. Experiência: arquivo de imagem contendo um dígito manualmente escrito e o valor do dígito nele contido. Medida de performance: a proporção de dígitos corretamente identificados.

Identificação de autoria de textos

O problema de identificação de autoria de textos consiste na identificação automatizada do autor de um texto.

Tarefa: dado um texto, identificar qual seu autor. Experiência: texto e respectivo autor. Medida de performance: a proporção de autores corretamente identificados.

problemas com tarefa, medida de performance, experiencia menos comuns

[9] [4]

bla bla bla de machine learning

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.

tabela com dados dos estudos analisados - ver fichamento.xlsx

descrever, em termos gerais, cada estudo

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

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.