Técnicas de Aprendizado de Máquina Aplicadas na Previsão de Evasão Acadêmica

Maurício J.V.Amorim[1], Dante Barone[2], André Uebe Mansur[1]

[1]Instituto de Informática – Centro Federal de Educação Tecnológica de Campos
Av.Dr.Siqueira,273 – Campos dos Goytacazes – RJ – Brazil

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{amorim,auebe}@cefetcampos.br, barone@inf.furb.br

Abstract *This article demonstrates the efficiency of the use of the machine learning techniques applied to the forecast of academic drop-out. It concentrates on the modeling of the main aspects that can make a student that hasn't finished his/her course yet, to cancel it or simply to abandon it .The article demonstrates the main phases for implementation of a forecast system (attributes selection, data research, classifiers choice), tests the accuracy of three classifiers widely used and display s the statistics regarding drop-out in each course.*

Resumo: *Este artigo demonstra a eficiência do uso das técnicas de aprendizado de máquina aplicadas à previsão de evasão acadêmica. Ele concentra-se na modelagem dos principais aspectos que possam levar um aluno que ainda não concluiu o seu curso, a trancá-lo ou simplesmente abandoná-lo. O artigo demonstra as principais fases para implementação de um sistema de previsão (seleção de atributos, levantamento dos dados, escolha dos classificadores), testa a acurácia de três classificadores amplamente utilizados e mostra as estatísticas referentes a evasão em cada curso.*

Palavras Chaves*: Aprendizado de Máquina, Máquinas de Vetores de Suporte, Árvores de Decisão, Sistemas Acadêmicos, Inteligência Artificial, Mineração de Dados, Ambiente WEKA.*

1. Aprendizado de Máquina

A aprendizagem de máquina é um sub-campo da inteligência artificial dedicado ao desenvolvimento de algoritmos e técnicas que permitam ao computador aprender, isto é, que permitam ao computador aperfeiçoar seu desempenho em alguma tarefa. .

Algumas partes da aprendizagem de máquina estão intimamente ligadas à mineração de dados e estatística. Sua pesquisa foca nas propriedades dos métodos estatísticos, assim como sua complexidade computacional. Sua aplicação prática inclui o processamento de linguagem natural, motores de busca, diagnósticos médicos, bioinformática, reconhecimento de fala, reconhecimento de escrita, visão computacional e locomoção de robôs e sistemas de previsão.

Neste artigo, vamos utilizar o aprendizado de máquina para tentar entender os motivos de evasão universitária, ou seja, descobrir os motivos que levam os alunos universitários a não concluírem seus cursos.

2. Evasão e suas características

A evasão universitária vem se impondo, ao longo do tempo, como uma realidade cada vez mais extensiva no âmbito do ensino de graduação (Cunha, 2000). Tal constatação, porém, ainda que reafirmada por números alarmantes, não vem se mostrando com força o bastante para "tocar as universidades em suas raízes" (Moraes, 1986).

No Brasil, e mesmo em outros países, a tendência dos estudos sobre evasão, de um modo geral, é a de orientar-se pela proposta do dimensionamento ou quantificação da evasão, havendo poucos estudos que tratem, qualitativamente, a questão (Ribeiro, 1996).

Conter a evasão é uma tarefa árdua. Para isto, o primeiro passo é conhecermos os motivos que levam os alunos a desistirem de tentar concluir o curso. O primeiro passo a ser dado é conseguir uma base de dados relativamente grande contendo as informações necessárias para este levantamento. Essa massa deverá ser utilizada como base para que o computador consiga extrair atributos necessários para entender o porque os alunos se evadem.

3. Como acontece o aprendizado de máquina

Para que se possa aplicar o aprendizado de máquina, se faz necessária a existência de uma massa de treinamento e teste, com diversos atributos que julgamos significativos além dos resultados que esperamos para cada um dos dados informados. Em outras palavras, para que o computador possa aprender ele necessita dos dados e das respostas.

Um exemplo prático já bastante utilizado é o reconhecimento óptico de caracteres. Nele, apresenta-se uma matriz com (n x m) elementos, onde cada elemento representa um pixel pintado ou não em um dispositivo de entrada. Após um treinamento inicial, o computador pode vir a conseguir reconhecer caracteres que sejam parecidos aos apresentados anteriormente.

Outro exemplo bastante citado na bibliografia é o da previsão baseado em fatores climáticos. Neste caso dispõe-se de uma base de dados contendo as condições de tempo (visibilidade, temperatura, umidade relativa do ar e velocidade do vento) de jogos anteriores e resultado se houve ou não jogo naquele dia. De posse desta massa anterior de dados, podemos aplicar a esta base a um algoritmo de treinamento, para que, após treinado, ele possa prever se, com as condições climáticas atuais haverá ou não jogo amanhã.

Como nos dois exemplos anteriores, uma modelagem inicial se fez necessária. Nessa modelagem a seleção dos atributos de entrada, seu formato e faixa de valores, a modelagem do formato da saída e a escolha do classificador (algoritmo que fará o treinamento) se integram como passos para sua aplicação. Muitas pesquisas estão sendo realizadas e diversas famílias de algoritmos de aprendizagem são amplamente utilizadas, cada uma sendo indicada para determinados tipos de problemas. Veremos

mais adiante uma ferramenta (WEKA, 2008) que agrupa um grande número de algoritmos para aprendizagem de máquina.

Para que se possa testar a eficiência dos algoritmos de classificação, se faz necessária a existência de uma segunda massa de dados que deverá ser utilizada para medir o desempenho do classificador, ou seja, seu grau de acurácia.

Um fator que influi no resultado é a quantidade de informações (elementos) contidas na massa de testes, uma vez que, em muitos casos, as informações existentes não são amplas o suficiente para que se consiga um treinamento adequado. Para contornar este problema, existem técnicas como a "cross validation" que, a exemplo, promove através de iterações, um aumento das comparações entre os poucos elementos, promovendo um incremento do treinamento e levando a um resultado mais satisfatório (WITTEN, 2005).

4. Aprendizado de máquina aplicada a evasão escolar

Em todo processo de classificação ou previsão através de aprendizado de maquina observamos certos procedimentos. O levantamento dos dados para treinamento e testes, a seleção dos atributos e sua transformação, a escolha dos classificadores, a execução do treinamento e avaliação dos resultados obtidos são fases obrigatórias para o bom desempenho. Após treinado temos um sistema capaz de através de novos atributos inseridos em sua entrada, realizar a classificação, ou como em nosso exemplo, a previsão da saída para aqueles dados. Veremos abaixo como ocorre cada uma destas fases em nosso estudo de caso.

4.1 Levantamento da massa de dados

Conforme levantamos anteriormente, existe a necessidade de uma massa de dados de treinamento e outra para teste para poder treinar e medir o desempenho dos algoritmos classificadores. No caso em análise, utiliza-se uma massa de testes que cobria 10 semestres letivos de uma universidade particular (protegendo a identidade dos alunos).

Os dados que compõem a massa de teste foram obtidos de uma Instituição de Ensino Superior (IES) particular, de caráter confessional, no município de Campos dos Goytacazes, RJ e que possuía, até a data de elaboração dos teste, 04 cursos de graduação plena: Administração, Arquitetura, Engenharia de Produção, Fisioterapia e Pedagogia. A escolha da IES se deu pela disponibilidade da base de dados por estudos acadêmicos anteriormente feitos a qual se utilizou a mesma base de dados.

Verifica-se que em uma massa de testes composta por um universo de 8073 matriculas realizadas nos cursos da IES, 1765 (21,9%) são do curso de Administração, 1160 (14,4%) do curso de Engenharia de Produção, 2642 (32,7%) são do curso de Fisioterapia e 2506 (31,0%) são do curso de Pegagogia.

Nessa instituição, estes dados encontravam-se em base de dados relacional, mais especificamente em um SGBDR Mysql 5.1(MYSQL, 2008) que concentrava toda história acadêmica e financeira dos alunos da instituição.

A figura 1 mostra o diagrama de estrutura de dados com todas as informações contidas na base de dados referenciada.

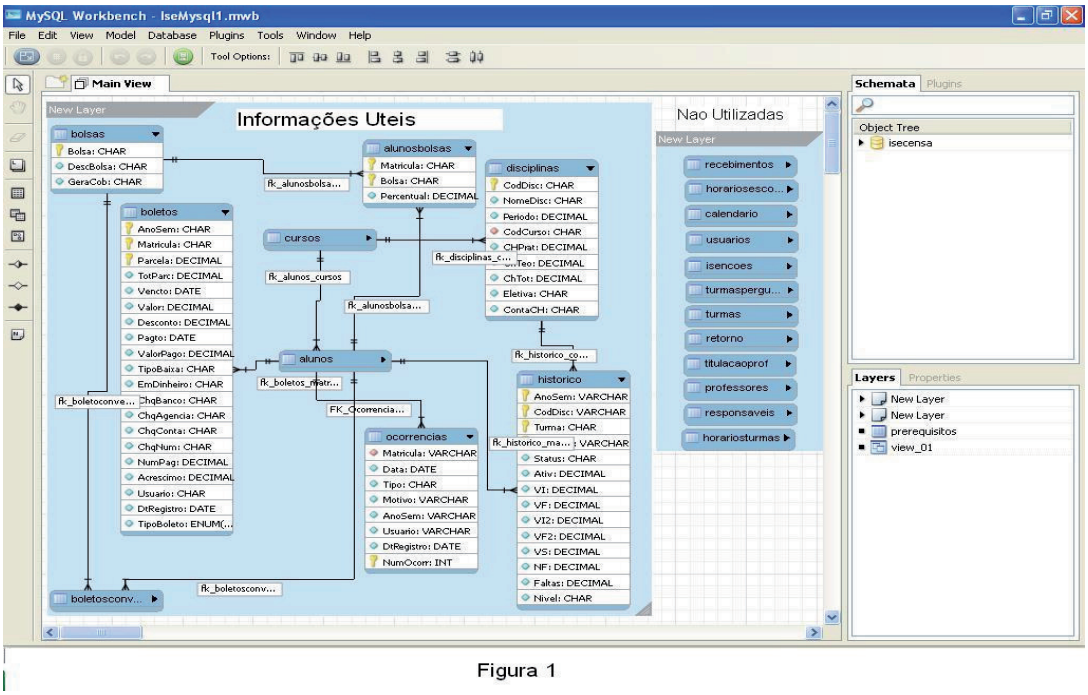

Figura 1 – DER dos dados da IES

Entre as informações que aparentemente úteis, pôde-se detacar :

- A matrícula de cada aluno (que foi modificada para proteger sua identidade);
- O ano e semestre de ingresso do aluno;
- A quantidade de disciplinas cursada pelo aluno no semestre anterior ao evento;
- O percentual de aprovação do aluno no semestre anterior, ou seja a relação entre o numero de disciplinas cursadas e o numero de disciplinas no qual ele foi aprovado;
- O percentual de desconto que o aluno possuía no semestre anterior ao evento;
- A quantidade de prestações em aberto que o aluno possuía no ato do evento, visto que o fator financeiro poderia vir a contribuir na evasão (em se tratrando de universidade particulares);
- Coeficiente de Rendimento Escolar (CR) do aluno, ou seja a média de todas as disciplinas que ele já cursou.
- A Quantidade de disciplinas do curso;
- O percentual do curso que o aluno já havia completado;

E por fim o principal atributo de treinamento:

- Eventos acadêmicos relevantes acontecidos durante a vida acadêmica do aluno. (renovações, trancamentos, cancelamentos, transferências ou conclusão);

Estas informações foram obtidas através da execução da consulta em sql sobre o banco de dados.

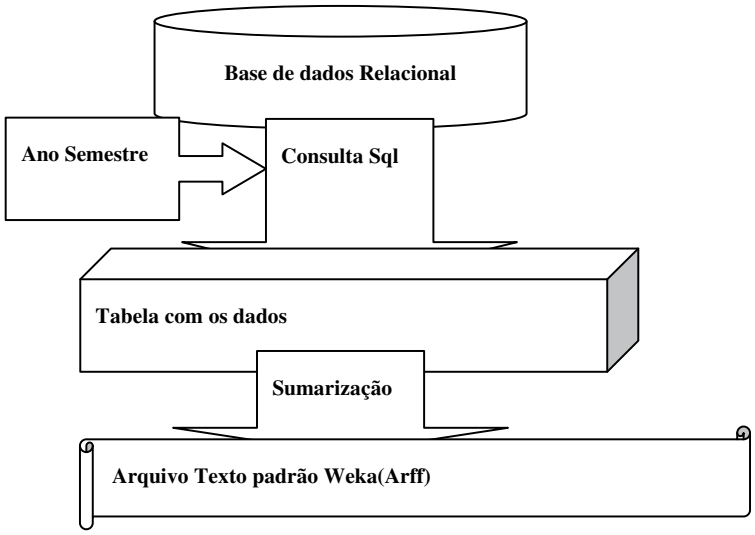

Figura 2. Arquitetura do Sistema Mapeador

4.2 Geração dos dados para os classificadores

Após a fase de pré-seleção dos atributos, com a massa de treinamento e testes disponível para uso, o passo seguinte é a escolha dos classificadores. Para facilitar esta tarefa, a utilização de uma ferramenta denominada Weka, construída pela Waikato University, na Nova Zelândia, veio contribuir imensamente (WEKA, 2008). Esta ferramenta consiste de um banco de programas em Java com os principais algoritmos e técnicas de *machine learning* disseminados, tais como : redes bayesianas, máquina de suporte de vetores, árvores de decisão, entre outras. Para que utilizá-la dever-se-ia transformar as massas de treinamento e a de testes em um formato adequado a esta ferramenta, ou seja o formato Arff. Para isto construí-se uma ferramenta em Delphi que obtinha os dados no SGBDR e transformava para o padrão requerido.

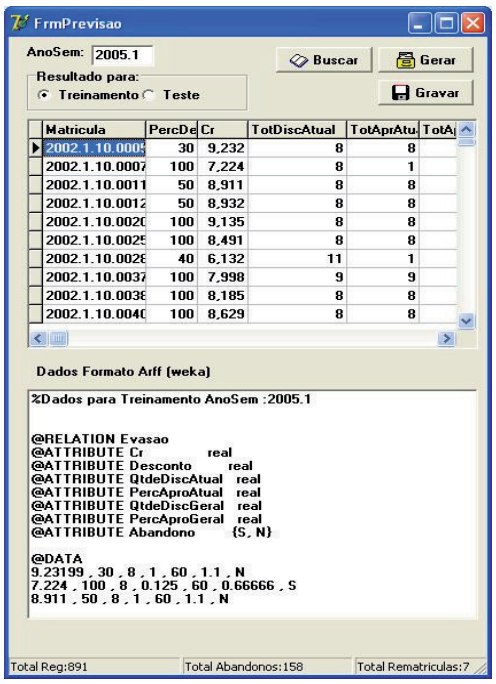

Figura 3. Tela do Sistema Mapeador

4.2 A escolha dos classificadores

Com os dados gerados no formato ARFF, o próximo passo consiste na escolha dos classificadores. Entre as dezenas de classificadores oferecidos pelo Weka, optou-se pela escolha de três classificadores:

- J48 – baseado em árvores de decisão;
- SMO – baseado em maquinas de vetores de suporte
- Bayes Net – baseados em métodos bayesianos;

Esta escolha baseou-se em orientação dada pela DSc. Bianca Zadrowzny, professora da cadeira de Aprendizado de Máquina, do Instituto de Computação da Universidade Federal Fluminense, a qual já desenvolve diversos trabalhos na área.

A estratégia empregada consistiu em treiná-los com os dados de um semestre anterior e testá-los com os dados do semestre seguinte. Partindo destes dados tinha-se um total de 3361 amostras, sendo 516 positivas ou seja, com evasão confirmada para treinamento. Para testes tinha-se um total de 8073 amostras sendo 1782 casos de evasão confirmados.

O manuseio do Weka é composto de 5 passos:

- Abertura dos arquivos de treinamento (figura 4);
- Escolha do classificador;
- Escolha dos métodos de testes, no caso em análise escolheu-se por fornecer o arquivo de testes a parte;

- Execução do treinamento e verificação dos resultados de testes;
- Geração do programa classificador, que no caso do Weka é uma classe em java obtida através dos dados de treinamento.

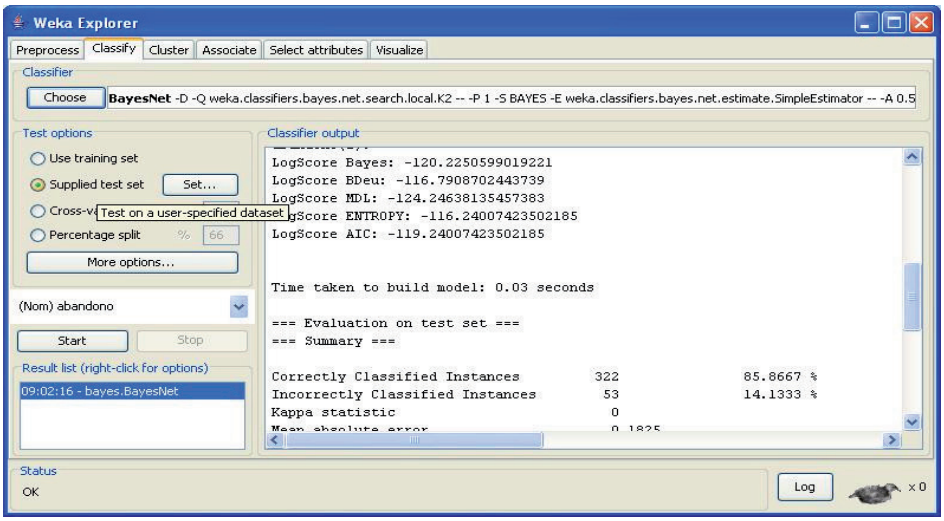

Figura 4. Tela do Weka

5. Resultados Computacionais

O resultado obtido com este levantamento gera duas grandes contribuições. A primeira contribuição é obtida através da análise do numero de evasões por curso. A segunda contribuição é a avaliação da eficácia de cada um dos algoritmos na previsão de evasão.

5.1 Evasão por Curso

Para entender os números apresentados sobre a evasão, levantamos o total de matriculas e re-matrículas ocorridas durante o período correspondente ao primeiro semestre de 2002 até o segundo semestre de 2006. Cada vez que um aluno novo entra no curso é computada uma matrícula ao curso e toda vez que um aluno antigo retornava no semestre seguinte é computado uma re-matrícula. Caso este aluno não retorne ao curso e não seja um aluno concludente, é computado um trancamento ao curso. A tabela 1 mostra os dados referentes aos cálculos apresentados.

Tabela 1 – Taxas de Evasão por curso

Cursos	Total de matrículas e re-matrículas	Total Trancamentos	Percentual Semestral de Evasão
Administração	1765	298	16,88%
Engenharia Produção	1160	363	31,29%
Fisioterapia	2642	558	21,12%
Pedagogia	2506	563	22,47%
Geral	8073	1782	22,07%

5.2 Eficácia de cada algoritmo na previsão da evasão

Lembramos que uma das finalidades do projeto é a construção de um sistema de previsão de evasão escolar, ou seja, construir um classificador já treinado, que computadas as entradas para os alunos em curso, ele possa indicar quais irão evadir-se, fazendo com que a IES possa tomar as devidas providências para tentar reverter o caso.

Para esta finalidade, ou seja, construir um sistema de previsão, se faz necessário submeter as bases de testes a diversos algoritmos de forma a avaliar o seu desempenho. Em outras palavras, necessita-se testar quais são os classificadores que obtém os melhores desempenhos nos acertos. Diante deste quadro, obteve-se os seguintes números (vide tabela 2). Os números obtidos foram bastantes próximos, com uma pequena vantagem para o SMO.

Tabela 2 – Grau de acurácia dos classificadores na evasão

	Bayes Net	SMO	J48
Classificacao Correta	89,7084%	91,2521%	89,6512%
Classificacao Incorreta	10,2916%	8,7479%	10,3488%

6. Considerações Finais

Este etudo apresenta uma ferramenta de geração de dados para classificadores padrão Weka, baseada em informações históricas obtidas através dos registros das informações acadêmicas do alunos de instituições de ensino particulares.

O artigo compara o desempenho de três classificadores amplamente utilizados (Bayes Net, SMO e J48) de a testar quais podem apresentar os melhores resultados obtidos para o problema em questão.

Tem-se, assim, uma grande contribuição do artigo como sendo a demonstração da utilidade das técnicas de *machine learning* no estudo da evasão, servindo de importante ferramenta para a percepção e tomada de decisões estratégicas no tratamento de problemas relacionados a evasão escolar.

Um passo importante que deve ser investigado em trabalho futuro é descobrir o peso de cada um dos atributos no processo de evasão e, para isto, sugere-se a utilização de métodos de análise qualitativa como o ELECTRE. (ROY, 1985, VINCKY, 1992)

Referências

BANDURA Bandura, A. (1977). **Self-efficacy: Toward a unifying theory of behavior change**. *Psychological Review, 84*, 191-215.

BANDURA, A. (1997). **Self-efficacy: The exercise of control**. New York: Freeman.

BICA, F., Verdin R. (2008). **InteliWeb: The E-Learning System that Recognizes Aspects of Self-Efficacy** em …

BRASIL. Ministério da Educação. **Lei de Diretrizes e Bases da Educação**. Disponível em: <www.mec.gov.br>. Acesso em: 10 ago. 2008.

CUNHA, A. et al (2000) **"Evasão do curso de química da Universidade de Brasília**: a interpretação do aluno evadido" - Química Nova -ISSN 0100-4042

MORAES, N. I. (1986) **"Perfil da universidade"**. São Paulo: Pioneira/Universidade de São Paulo.

RIBEIRO, C. (1996) **"As causas da evasão universitária"** em Anais do I Encontro Setorial dos cursos de Graduação da UNESP 1996, 176.

MYSQL Disponível em http://www.mysql.com/documentation/. Acesso em 10 ago. 2008.

RUSSEL, S., NORVIG,P. (2004). **Artificial Intelligence: A Modern Approach.**, 2nd ed. Editora Campos.

WEKA – Disponível em http://weka.sourceforge.net/wekadoc/. Acesso em 10 ago. 2008.

WITTEN, I.H. and FRANK, E. (2005) **Data Mining: Practical machine learning tools and techniques**. *In 2nd edition Morgan Kaufmann*, San Francisco.

VINCKE Ph.(1992) **Multicriteria decision-aid**, Wiley, New York, 1992.