# Página 1

Previsão do abandono do comportamento social dos alunos

## Jaroslav Bayer, Hana Bydzovská, Jan Géryk, Tomáš Obšıvac

Unidade de Sistemas de Computador Faculdade de Informática, Universidade Masaryk

Brno, República Tcheca

## { bayer, bydzovska, geryk, obsivac } @ fi.muni.cz

Lubomır Popelınský

Grupo de Descoberta de Conhecimento Faculdade de Informática, Universidade Masaryk

Brno, República Tcheca

## [popel@fi.muni.cz](mailto:popel@fi.muni.cz)

RESUMO

Este artigo se concentra na previsão de evasão e reprovação escolar

ures quando os dados do aluno foram enriquecidos com dados derivados do comportamento social dos alunos. Esses dados descrevem dependências comerciais coletadas de e-mail e fórum de discussão

conversas, entre outras fontes. Descrevemos um extrac-

ção de novos recursos de dados e comportamento dos alunos dados representados por um gráfico social que construímos. Então

apresentamos um novo método para aprender um classificador para estudos

previsão de falha dentada que emprega aprendizagem sensível ao custo para diminuir o número de malsucedidos classificados incorretamente alunos. Mostramos que o uso de dados de comportamento social resulta em aumento significativo da precisão da previsão.

como intensidade de comunicação interpessoal ou número de

arquivos mutuamente compartilhados, podem ser observados e armazenados imediatamente, quando a função particular do sistema é usada,

ou mais tarde a partir do histórico completo de solicitações de usuários que está presente na forma de registro de acesso ao sistema. Relações

entre os alunos (identificados a partir de seu comportamento social) são principais blocos de construção de uma rede social latente. Com o

ajuda da Análise de Redes Sociais (SNA) [4], calculamos vários vários novos recursos de um aluno da rede, por exemplo

características dos vizinhos.

Neste artigo, apresentamos um novo método para geração de dados ação, pré-processamento e mineração de dados educacionais (EDM) [1; 14; 10] que utilizam os registros do aluno e os dados

## INTRODUÇÃO

Uma das tendências atuais no ensino superior é o sub

aumento substancial de alunos do primeiro ano e, consequentemente, o volume de dados educacionais. Milhares de alunos são

admitido a estudar em universidades todos os anos. Eles alcançam

resultados finais, aprovação ou reprovação nos exames, comunique-se com cada um outros durante seus estudos e muitos deles não conseguem terminar

seu estudo com sucesso. Os funcionários da universidade gostariam de entrar

coragem esses alunos para terminar seus estudos, mas é difícil para identificá-los cedo também devido ao grande número de alunos matriculados. É importante explorar métodos que

pode extrair conhecimento confiável e abrangente do

dados do aluno que permitem a previsão de um abandono com um suf- precisão suficientemente alta.

Neste trabalho, utilizamos dados de alunos que foram armazenados no Sistema de Informação da Masaryk University (IS MU),

que armazena dados educacionais e compreende todas as informações informações sobre alunos e seus estudos, sobre professores e

cursos, e também fornece ferramentas de gerenciamento de exame, sistema de registro de desculpas, avaliação de testes on-line e várias formas de comunicação, por exemplo, fóruns de discussão.

Utilizamos apenas um subconjunto de informações armazenadas no IS MU que é relevante para a previsão do sucesso do aluno, como

notas de teste de capacidade de estudo, créditos ganhos, notas médias,

ou gênero. Os dados do IS MU são importados periodicamente para data warehouse Excalibur [3] que combina três dis-

disciplinas de processamento de dados - gerenciamento de dados, mineração de dados (DM) e análise visual.

O IS MU também armazena o histórico completo das solicitações dos usuários para o sistema. Dados sobre o comportamento social dos alunos, como

sobre seu comportamento social. Mostramos como prever estu-

evitar o abandono escolar e o fracasso escolar usando métodos DM [7] e

SNA. Usamos SNA para criar novos recursos relacionados ao estudo

que pode ajudar os métodos convencionais de aprendizagem a aumentar o precisão de prever o desempenho do aluno ou detectar um

possível desistência. Pretendemos construir classificadores para os primeiros

detecção e previsão de longo prazo de um potencial abandono. A detecção precoce implica a necessidade de histórico de dados.

Os resultados preliminares para esta tarefa foram publicados no doutorado workshop [2]. A maior precisão medida foi acima

80% quando apenas os dados dos alunos foram empregados. Nós enriquecemos os dados do aluno com os dados sobre o comportamento social e

alcançou um aumento da precisão geral de cerca de 10%.

Em ambos os casos, o aprendizado de máquina baseado no ganho de informações (ML) métodos geraram os classificadores de maior sucesso.

Outra abordagem para a previsão de um estudo do aluno por- desempenho que é baseado em questionários pode ser encontrado em [12]. Em [15], o projeto de um sistema baseado na web para resolver questões relacionadas ao desempenho do aluno no ensino superior

é proposto. Ele utiliza uma implantação de função de qualidade em

combinação com métodos de DM. Um novo método de ML pré-

eliminando o abandono do ensino superior a distância por causa de desequilíbrios conjuntos de dados anced são discutidos em [9]. Revela limitações de

os métodos existentes e propõe outra abordagem baseada

em técnicas sensíveis ao custo local. Uma nova abordagem para identificar todos os fatores que influenciam o sucesso do aluno são discutidos em [11]. Ele se concentra em fatores disponíveis antes do início de

um programa de graduação de alunos sugerindo regras associativas para descoberta de subgrupo para prever possíveis desistências. Um significante não melhora a previsão de abandono de calouros usando

a aprendizagem sensível ao custo é descrita em [5]. A maior acu

A rapidez da classificação foi alcançada usando árvores de decisão. No

comparação com a nossa abordagem de utilização do comportamento social, uma combinação de métodos de mineração de dados com LAN natural

Anais da 5ª Conferência Internacional de Mineração de Dados Educacionais 103

# Página 2

processamento de calibre, especialmente mineração de texto, foi empregado em

[17] para aumentar a retenção de alunos.

Na seção a seguir, apresentamos a estrutura de ambos

os dados do aluno e os dados de comportamento social e o nec-

etapas de pré-processamento essenciais. Descrevemos como construímos o rede social e aplicou os métodos analíticos na Seção

2.2. Seção 2.3 descreve o método DM usado para abandono predição. Na Seção 3, demonstramos os resultados e o melhoria da classificação medindo a quantidade

dos dados adicionais explorados pelo SNA. Então nós mostramos que classificadores de alta precisão podem ser criados para cada aluno independentemente da fase real do estudo. Discussão de re-

resultados está na Seção 4. Finalmente, concluímos este artigo com um visão geral dos principais resultados e trabalhos futuros na Seção 5.

## PREVISÃO DE DADOS E DROP-OUT

* 1. Dados do aluno

Nossa pesquisa considera alunos de bacharelado da Applied Infor- matemática admitida na Faculdade de Informática, Masaryk Univer- cidade nos anos de 2006, 2007 e 2008. Para esse período, podemos obter dados que correspondam a todo o comprimento do padrão

estudo de bacharelado, ou seja, três anos. O ano de 2006 como o menor limite é definido como o ano em que os dados de comportamento social começou a coletar. Exploramos apenas os alunos que

estiveram em contato com a comunidade escolar. Tais alunos produzir dados de comportamento social caracterizando-os no ambiente universitário.

Selecionamos apenas atributos gerais de estudos para ser capaz de

aplicar nossa abordagem a alunos de qualquer corpo docente. Para prever um abandono durante todo o período do estudo que coletamos

instantâneos de dados para cada período de estudos do aluno. O conjunto de atributos podem ser divididos em três categorias de acordo

para o tipo: atributos relacionados ao aluno, relacionados ao semestre em-

tributos e atributos relacionados a outros estudos.

Os atributos relacionados ao aluno incluem o seguinte:

(1) gênero

1. cursos incompletos - o número de cursos de um aluno dente falhou em completar
2. segundo resit feito - o número do segundo utilizado resits. Cada aluno pode exercer o direito ao segundo

repor apenas tantas vezes quanto o comprimento padrão de o estudo em anos aumentou em um.

1. dias dispensados - o número de dias em que um aluno é expulso
2. notas médias - a nota média calculada de todos ganhou notas
3. notas médias ponderadas - notas médias ponderadas pelo número de créditos ganhos para os cursos
4. a proporção do número de créditos ganhos para o número de créditos para ganhar
5. a diferença de créditos ganhos e créditos para ganho

Porque um aluno pode estar matriculado em mais estudos ou também

em mais faculdades, adicionamos também atributos relacionados a outros estudos do aluno. Este conjunto de atributos consiste no

Segue:

1. o número de estudos paralelos na faculdade
2. o número de estudos paralelos na universidade
3. o número de todos os estudos na faculdade
4. o número de todos os estudos na universidade

Dados que consistem em valores de todos os atributos que caracterizam um estudo em um ponto do tempo foi extraído de Excal-

ibur. O conjunto de dados continha 775 alunos, 837 estudos e

* 1. ano de nascimento
  2. ano de admissão
  3. isenção de vestibular
  4. pontuação no teste de capacidade de estudo - um resultado da entrada exame expresso como a porcentagem da pontuação

medindo potencial de aprendizagem

Os atributos relacionados ao semestre são os seguintes:

* 1. o número de semestres concluídos
  2. cursos reconhecidos - o número de cursos relacionados terminou em outros estudos
  3. créditos reconhecidos - o número de créditos ganhos de cursos reconhecidos
  4. créditos a ganhar - o número de créditos a ganhar por en- cursos rolados, mas ainda não terminados
  5. ganhou créditos - o número de créditos ganhos de fin- cursos ished

4.373 exemplos no total - um exemplo por termo, onde o o número de termos para um aluno variou de 1 a 8.

## Dados de comportamento social

O conjunto acima mencionado de 775 alunos é o núcleo do ego rede social centrada. Nós o criamos a partir dos alunos e

seus colegas de escola diretos e relações entre eles. Rela-

ções refletem os padrões de dados de comportamento social. Então nós computar novos recursos do aluno a partir da estrutura da rede

características e atributos de vizinhos diretos do aluno.

Para obter conhecimento sobre um aluno a partir da perspectiva de seu envolvimento na comunidade escolar, nós con-

struct um sociograma, um diagrama que mapeia a estrutura de relações interpessoais. Esse gráfico social permite encontrar novos recursos por classificação baseada em link.

Existem vários laços interpessoais já avaliados para

aprimorar a pesquisa de texto completo do IS MU. Nós os computamos em linha ou através do processamento de registro do sistema e armazená-los em o índice do motor de busca como um documento relevante não textual tokens e como parte do modelo do usuário. Estes são então

usado para melhor ordenar os resultados da pesquisa por correspondência de documentos

mentos (por exemplo, e-mails, arquivos, cursos) relacionados aos respectivos usuários [16].

Anais da 5ª Conferência Internacional de Mineração de Dados Educacionais 104

# Página 3

* + 1. Nova geração de recursos

Esta rede social de modo único de alunos e seus inter-

laços pessoais (ou seja, rede de informação homogênea) permite

para explorá-lo não apenas visualmente, mas também por meio de ferramentas sociais análise de rede, por exemplo, Pajek [13]. Além disso, anteriormente

Figura 1: Rede com vértices organizados por Kamada-Kawai

Pajek

características vistas de cada aluno podem ser calculadas com tais Ferramentas. Os dois tipos de recursos a seguir são interessantes e nos dar uma nova visão dos dados.

Em primeiro lugar, os recursos obtidos a partir da estrutura de rede são com- colocado a partir de características estruturais básicas, ou seja, a versão

algoritmo de layout de energia. Nós escuros representam alunos

com estudos concluídos com sucesso.

Alguns laços são fatos intuitivos e fortes, a saber:

1. amizade explicitamente expressa
2. conversa mútua por e-mail
3. publicação em coautoria
4. comentário direto sobre outra pessoa

Os laços mais fracos são mais ocultos e são derivados do seguinte fatos a seguir:

1. mensagem do fórum de discussão marcada como importante
2. tópico inteiro no fórum de discussão ou blog marcado como fa- vourite
3. arquivos enviados para o depósito de outra pessoa
4. avaliações das mensagens do quadro de avisos
5. visitou páginas pessoais

Medimos o valor de um empate por sua importância e peso

pelo número de ocorrências. Por exemplo, um empate representa envio de troca de vários e-mails tem valor maior do que

uma visita ao perfil pessoal de alguém. A identificação de

os melhores pesos é um assunto possível para avaliação futura. Outra propriedade notável de uma relação é sua direção. É em dica a origem e o alvo de uma ação que contamos

como a relação. Por exemplo, uma pessoa que enviou / recebeu um e-mail ou quem carregou / recebeu um arquivo na fonte / destino

respectivamente. Algumas ações não têm direção, por exemplo, marcação

o mesmo tópico de discussão que um favorito.

Como resultado, calculamos um único número de todos os mencionados laços refletindo a força geral de uma relação do aluno com

graus tex, o resumo dos valores da linha incidente e o

centralidade de intermediação:

1. grau - o número de linhas que incidem com um ver- tex, representa quantas relações o aluno está em envolvido em
2. indegree (ou popularidade) - o número de arcos vindo para o nó, ele representa quantos outros membros da rede o aluno é um assunto de interesse
3. outdegree - o número de arcos com direção oposta

representa um interesse iniciado por um determinado aluno

1. soma dos valores da linha incidente - para medir também o força dos laços
2. centralidade de intermediação - o número de caminhos mais curtos de todos os vértices para todos os outros que passam por dados vértice representa a importância do aluno (global para o

rede)

Em segundo lugar, os recursos obtidos da própria vizinhança laços também são importantes para examinar, e devemos medir

não apenas a quantidade de vínculos pessoais, mas também sua qualidade.

Em outras palavras, o desempenho acadêmico do entorno

estudar os alunos é importante, porque seria difícil conseguir vantagem da comunicação com alunos malsucedidos.

Selecionamos quatro recursos de alunos do conjunto de dados,

transferido por seu ganho de informação, para calcular as médias do valores de vizinhança (ANV):

1. pontuação do teste de capacidade de estudo ANV
2. ANV média da nota
3. proporção de créditos inscritos e cumpridos ANV
4. créditos por semestre ANV 1
   1. Processo de previsão de abandono

Nosso objetivo era desenvolver um método preciso para abandono

qualquer colega de escola. Encontramos 13.286 dessas conexões representando arcos de gráfico (linhas orientadas) avaliados por este número

ber.

Agora, a rede pode ser visualizada para análises exploratórias

sis de suas propriedades. Por exemplo, depois de aplicar Kamada- Algoritmo de layout de energia Kawai [8] (Fig. 1), podemos ver que os alunos bem-sucedidos (nós negros) ocupam a área no

meio da rede e raramente são vistos na periferia.

Ao contrário, os malsucedidos (nós brancos) são colocados

em todo o gráfico. Isso, junto com os resultados apresentados posteriormente, apóia nossa suposição de que um número maior e mais forte

os laços têm um impacto positivo no sucesso do estudo, enquanto a ausência dos empates prevê um potencial de fracasso.

previsão que também permitiria prever o abandono um estágio inicial do estudo. O método deve ter min-

número de falsos negativos, ou seja, alunos que não foram reconhecido como estando em perigo de abandono escolar. Quando todos os atributos foram usados, a precisão foi fraca.

É por isso que utilizamos métodos de seleção de recursos para reduzir

a dimensionalidade dos dados do aluno extraídos do Ex-

calibur data warehouse. Melhoramos o pré-processamento método descrito em [12], calculando a classificação média de atributos enquanto elimina os valores extremos.

1 Surpreendentemente, quando tentamos usar esses recursos com ponderação usando a força da conexão correspondente

ção, não melhorou o desempenho dos classificadores.

Anais da 5ª Conferência Internacional de Mineração de Dados Educacionais 105

# Página 4

O objetivo era preservar a confiabilidade dos atributos para classificação catião após a redução. Portanto, utilizamos uma combinação

ção de algoritmos de seleção / estimativa de recursos com base em dife- abordagens diferentes. Empregamos três algoritmos baseados em

entropia (InfoGainAttributeEval, GainRatioAttributeEval e SymmetricalUncertAttributeEval), um algoritmo select- o atributo de erro mínimo para previsão (OneRAt-

tributeEval), um algoritmo que utiliza a distribuição χ 2 (Chi-

SquaredAttributeEval), um algoritmo que prefere atributos

altamente correlacionado com a classe, mas com baixa intercorrela- para outros (CfsSubsetEval), um algoritmo que procura o

menor subconjunto de atributos com consistência igual ao de todos os atributos (ConsistencySubsetEval), e um

algoritmo avaliando atributos encontrando o vizinho mais próximo

bours para um exemplo escolhido aleatoriamente de cada classe. Isto compara as diferenças acumuladas de valores do cor-

comportamento social não superou 90% e o melhor resultado foi obtido com o aluno da árvore de decisão, 82,53%, e o TP taxa de 78,50%.

Os atributos mais significativos incluem a proporção do número número de créditos ganhos com o número de créditos a ganhar, e a média desta relação medida para vizinhos ponderada

pela força de sua relação na rede social. o

sete atributos mais relevantes são apresentados na Tabela 1.

Tabela 1: Sete os atributos mais relevantes Média do pedido Ord. Atributo

1 1,000 (16)

|  |  |  |
| --- | --- | --- |
| 2 | 2.000 | (14) |
| 3 | 2.625 | (15) |

recursos de resposta (ReliefFAttributeEval), e utilizamos também dois filtros (FilteredAttributeEval, FilteredSubsetEval).

Em seguida, calculamos uma lista de atributos ordenados pelo av- graus erage obtidos a partir das listas ordenadas produzidas pelo

algoritmos de seleção de recursos avaliando a importância de os atributos. Para cada atributo, pulamos o extremo

valores - as melhores e as piores avaliações. Nós reduzimos o

conjunto de atributos para os 22 mais relevantes e aprenderam a

classificadores novamente. Exceto para o método Naive Bayes (NB),

todos os métodos de aprendizado de máquina usados obtiveram uma maior precisão atrevido. Exemplos dos atributos removidos são os seguintes:

ser um tutor do seminário, o número de alterações de senha, ou o número de cursos inscritos.

A lista do conjunto refinado de atributos em ordem de relevância pode ser encontrado na Tabela 1.

Em seguida, calculamos características estruturais significativas de a rede social para obter atributos adicionais que impliquem relações sociais entre os alunos.

Empregamos métodos de aprendizado de máquina de Weka no dados do aluno e, em seguida, nos dados que também continham o

dados de comportamento social. Para cobrir todos os tipos de aprendizado de máquina algoritmos de processamento, empregamos o aprendiz da árvore de decisão J48, IB1

aluno preguiçoso, aluno de regras PART, vetor de suporte SMO ma-

chines e classificador NB. Também empregamos o aprendizado de conjunto métodos de seleção, nomeadamente ensacamento e votação. Utilizamos custo-

aprendizagem sensível (CSM) e, em seguida, ensacamento com matriz de custos. Todos os métodos foram usados com configurações de parâmetro padrão.

O desempenho foi medido em termos de precisão (o número

número de exemplos classificados corretamente em relação ao número de todos exemplos) e taxa de verdadeiro positivo (o número de

exemplos classificados da classe de alunos malsucedidos). Usamos validação cruzada de 10 vezes.

## RESULTADOS

Primeiro, criamos um classificador usando apenas o comportamento social dados, mas a precisão não subiu acima de 69%, na verdade,

foi menor do que para aprender com os dados dos alunos. No entanto, se

adicionamos os atributos que descrevem o comportamento social aos dados do aluno, observamos um aumento de precisão

que chegou a 11%. Os principais resultados podem ser encontrados na Tabela 2. Em

a primeira coluna representa os resultados obtidos do Excal- ibur data warehouse, seguido pelos resultados do Excal-

Dados ibur enriquecidos pelos dados de comportamento social. A base- linha era 58,86%. A maior precisão foi obtida com

4 4.500 (5)

5 5,625 (17)

6 6.000 (8)

7 7.750 (10)

Tabela 2: Aprendizagem com os dados dos alunos (Excalibur) e alunos dados dentados enriquecidos com atributos de comportamento social (com SNA) [%]

Excalibur Com SNA

Método preciso. TP Preciso TP

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ZeroR | 58,86 | - | 58,86 | - |
| NB | 77,57 | 73,5 | 72,26 | 83,4 |
| SMO | 79,17 | 64,6 | 81,59 | 74,2 |
| IB1 | 78,14 | 72,5 | 89,80 | 86,2 |
| PAPEL | 82,44 | 73,7 | 93,67 | 92,3 |
| OneR | 75,89 | 57,9 | 88,45 | 83,8 |
| J48 | 82,53 | 78,5 | 89,89 | 88,8 |

Consideramos os dados de comportamento social uma característica de um estudante. Portanto, aprendemos classificadores apenas com o dados de comportamento social sem instantâneos de estudos de alunos dados. A linha de base foi ligeiramente mais baixa do que para o aluno dados ou os dados enriquecidos. O classificador de maior sucesso foi PARTE com a precisão de 68,82% e o TP com a taxa

70,50%. Os resultados estão na Tabela 3.

Tabela 3: Aprendendo apenas com atributos de comportamento social [%]

Método preciso. TP

ZeroR 50,18 -

|  |  |  |
| --- | --- | --- |
| NB | 64,04 | 80,6 |
| SMO | 63,68 | 83,5 |
| IB1 | 60,10 | 63,5 |
| PAPEL | 68,82 | 70,5 |
| OneR | 59,50 | 57,3 |
| J48 | 68,34 | 65,0 |

Em seguida, analisamos o sucesso de uma previsão de abandono

seria por diferentes períodos de tempo. Nós aprendemos classificadores

em resultados de estudos provisórios enriquecidos por dados de comportamento social

PARTE, 93,67%, e a taxa de Verdadeiro Positivo (TP) 92,30%. Precisão para os dados sem informações sobre o aluno

para reconhecer desistências o mais rápido possível. Resultados em termos de precisão (%) estão na Tabela 4.

Anais da 5ª Conferência Internacional de Mineração de Dados Educacionais 106

# Página 5

Tabela 4: Aprendizagem de dados de alunos enriquecidos com atributos de comportamento social por semestre [%]

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | 1 | 2 | | 3 | | 4 | | 5 | | 6 | | 7 + | | |
| Método preciso. TP | |  | Preciso TP | | Preciso TP | | Preciso TP | | Preciso TP | | Preciso TP | | Preciso TP | | |
| ZeroR | 50,18 | - | | 50,25 | - | 53,87 | - | 58,56 | - | 64,02 | - | 72,20 | - | 76,77 | - |
| NB | 71,45 | 69,1 | | 78,87 | 75,8 | 78,98 | 80,7 | 78,77 | 81,8 | 78,66 | 80,2 | 77,56 | 76,3 | 68,60 | 68,0 |
| SMO | 72,40 | 73,9 | | 81,33 | 80,2 | 81,02 | 77,5 | 83,22 | 78,1 | 83,74 | 72,3 | 87,56 | 67,5 | 85,48 | 52,3 |
| IB1 | 66,48 | 62,4 | | 70,64 | 67,2 | 66,72 | 61,1 | 71,40 | 63,2 | 74,59 | 61,0 | 77,07 | 53,5 | 90,93 | 75,8 |
| OneR | 62,84 | 65,7 | | 77,89 | 77,3 | 79,71 | 74,4 | 83,56 | 74,4 | 81,50 | 66,7 | 83,90 | 60,5 | 80,58 | 37,5 |
| PAPEL | 70,13 | 69,5 | | 74,82 | 74,3 | 76,20 | 72,8 | 76,20 | 73,1 | 77,24 | 69,5 | 79,51 | 64,0 | 91,11 | 83,6 |
| J48 | 70,73 | 71,2 | | 74,82 | 72,8 | 75,77 | 72,5 | 77,91 | 72,7 | 77,64 | 67,8 | 80,00 | 63,2 | 87,11 | 68,8 |

Posteriormente, nos concentramos na previsão de desistências quando a história dos dados sobre os estudos dos alunos é empregada. Tudo

instantâneos de dados foram usados. Resultados em termos de precisão (%)

estão na Fig. 2. No eixo X, há um período de estudo em

semestres (por exemplo, 3 significa que apenas os dados dos 3 primeiros semestres foram usados para construir o classificador). Mais

os detalhes estão na Tabela 5 e na Tabela 6.

Tabela 7: Precisões de meta-classificadores

|  |  |  |
| --- | --- | --- |
| Preciso | | TP ICUS |
| Excalibur (J48) CSM 80,45 85,7 258 | | |
| Com SNA (PART) CSM | 92,89 92,8 | 129 |
| Excalibur (J48) Bagging | 83,30 87,8 | 219 |
| Com SNA (PART) Bagging | 96,66 96,0 | 55 |

citar alunos com boas notas pode ser bem-sucedido

se formar com uma probabilidade maior do que os alunos com similar desempenho lar, mas não se comunicando com os alunos de sucesso dentes. Identificamos instâncias classificadas erroneamente e sup-

Figura 2: Classificações de acordo com os semestres

Podemos ver que para todos os períodos a classificação que utilizou apenas os dados do aluno alcançam menor precisão em comparação à classificação nos dados enriquecidos. Além disso, comece

combinando com o período dos primeiros quatro semestres a precisão de classificação nos dados enriquecidos foi superior a 90%.

Podemos concluir que quatro semestres é um período em que nosso modelo pode prever uma desistência com alta probabilidade. Nós con considerar este resultado satisfatório. The Masaryk University

avalia o potencial de aprendizagem dos alunos antes que eles sejam admitido para estudar.

Para a nossa tarefa é mais sério quando um aluno não é reconhecido considerado em perigo de abandono do que a situação oposta

ação. Para diminuir o número de classificados incorretamente

alunos bem-sucedidos, testamos a aprendizagem sensível ao custo (CSM)

e também ensacamento e, em seguida, ensacamento com matriz de custo, al- maneiras com o algoritmo de aprendizagem mais preciso como base

classificador. No caso de aprendizagem sensível ao custo, definimos um custo matriz para [0, 1, 0,5, 0] de modo que o custo do erro falso negativo

(ou seja, de alunos fracos não reconhecidos) foi duas vezes maior.

Todos os resultados estão na Tabela 7 na forma de Precisão (%),

Taxa de TP (%) e estudos malsucedidos classificados incorretamente (ICUS).

## DISCUSSÃO

Com base nos resultados, concluímos que o desempenho de um aluno mance parece estar correlacionada com os hábitos sociais, principalmente com a frequência da comunicação. Suporta o hy-

hipótese de que os alunos com resultados médios, mas comunicam

complementou-os com informações adicionais sobre

cursos. Descobrimos que cerca de um terço dos alunos não

completar dois cursos particulares (autômatos e gramáticas

e inglês especializado). Essas descobertas podem ser úteis em o trabalho futuro.

Classificadores com base no ganho de informação foram os mais bem sucedidos cessantes. O classificador NB sofreu com a forte in-

pressuposto de dependência, em nossos dados.

Também combinamos os dois classificadores de maior sucesso - J48 e PART - e construiu um meta-classificador onde a previsão

foi calculado como a média das probabilidades de

classificadores. No entanto, a precisão geral não foi maior do que o do melhor classificador.

Nós investigamos a influência dos dados de comportamento social em a precisão da classificação com relação ao gênero de

alunos. Os dados adicionais não aumentaram a precisão

em absoluto. Qualquer classificador não superou a linha de base de 92,11%. Em comparação com [12], empregamos a análise de redes sociais.

Eles alcançaram maior precisão, mas com at-

homenagens obtidas a partir dos dados coletados especialmente para o estudo. Esses atributos não podem ser recuperados de

sistemas de informação escolar padrão, por exemplo, hábitos de fumar, o nível de educação dos pais ou o número de irmãos.

Nós investigamos a influência do aprendizado sensível ao custo em a precisão de uma previsão de abandono. Empregando um custo matriz não diminuiu a precisão geral, mas ligeiramente

melhorou a taxa de TP. Usando ensacamento com uma matriz de custos aumentou a precisão e a taxa de TP. Dentro do estojo

de classificação nos dados do aluno, a precisão permaneceu quase inalterado, mas a taxa de TP aumentou de 78,5%

a 87%. A melhoria mais significativa foi alcançada em

o caso de classificação nos dados enriquecidos. O meta-

classificador aumentou a precisão para 96,66% e a taxa de TP

a 96%. O número de instantâneos de dados de classificação incorreta número de alunos malsucedidos diminuiu de 146 para 55 no

caso da classificação usando PART. O número de todos instantâneos de dados é 4.373.

Anais da 5ª Conferência Internacional de Mineração de Dados Educacionais 107

# Página 6

Tabela 5: Aprendendo com os dados dos alunos apenas de acordo com o semestre [%]

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | 1 | 1-2 | | 1-3 | | 1-4 | | 1-5 | | 1-6 | | Tudo | | |
| Método preciso. TP | |  | Preciso TP | | Preciso TP | | Preciso TP | | Preciso TP | | Preciso TP | | Preciso TP | | |
| ZeroR | 50,18 | - | | 50,21 | - | 51,28 | - | 52,74 | - | 54,37 | - | 56,28 | - | 58,86 | - |
| NB | 63,80 | 34,5 | | 70,56 | 50,5 | 72,47 | 55.0 | 74,66 | 59,1 | 75,82 | 67,4 | 76,64 | 72,7 | 77,57 | 73,5 |
| SMO | 69,41 | 64,7 | | 72,62 | 61,9 | 75,26 | 63,1 | 76,58 | 64,9 | 77,64 | 65,5 | 78,41 | 65,4 | 79,17 | 64,6 |
| IB1 | 62,72 | 61,2 | | 66,38 | 66,4 | 69,43 | 67,0 | 70,96 | 68,6 | 72,30 | 68,8 | 74,73 | 70,2 | 78,18 | 72,3 |
| OneR | 55,56 | 41,0 | | 64,93 | 68,1 | 70,63 | 76,5 | 74,14 | 79,1 | 75,32 | 76,0 | 75,27 | 70,9 | 75,90 | 57,9 |
| PAPEL | 65,35 | 73,4 | | 71,29 | 71,5 | 76,33 | 71,8 | 78,97 | 73,3 | 80,01 | 75,0 | 81,34 | 77,9 | 82,44 | 73,7 |
| J48 | 61,77 | 62,8 | | 71,77 | 73,0 | 75,47 | 73,6 | 77,67 | 75,2 | 79,34 | 75,5 | 80,61 | 77,1 | 82,53 | 78,5 |

Tabela 6: Aprendizagem de dados de alunos enriquecidos com atributos de comportamento social de acordo com o semestre [%]

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | 1 | 1-2 | | 1-3 | | 1-4 | | 1-5 | | 1-6 | | Tudo | | |
| Método preciso. TP | |  | Preciso TP | | Preciso TP | | Preciso TP | | Preciso TP | | Preciso TP | | Preciso TP | | |
| ZeroR | 50,18 | - | | 50,21 | - | 51,28 | - | 52,74 | - | 54,37 | - | 56,28 | - | 58,86 | - |
| NB | 71,45 | 69,1 | | 75,05 | 75,4 | 75,81 | 78,3 | 75,41 | 79,7 | 75,41 | 80,7 | 74,80 | 80,9 | 74,07 | 80,8 |
| SMO | 72,40 | 73,9 | | 77,10 | 75,7 | 79,15 | 76,7 | 80,10 | 77,5 | 80,36 | 76,4 | 81,66 | 76,7 | 81,68 | 74,4 |
| IB1 | 66,43 | 62,4 | | 67,41 | 63,7 | 70,59 | 67,4 | 76,92 | 73,1 | 81,07 | 76,8 | 83,10 | 79,2 | 90,10 | 86,7 |
| OneR | 62,84 | 65,7 | | 69,11 | 67,0 | 74,83 | 74,0 | 81,27 | 79,7 | 83,56 | 81,5 | 82,31 | 79,7 | 88,20 | 83,6 |
| PAPEL | 70,13 | 69,5 | | 79,65 | 77,6 | 86,60 | 86,7 | 90,21 | 89,3 | 92,38 | 90,9 | 92,99 | 91,1 | 93,51 | 91,9 |
| J48 | 70,73 | 71,2 | | 80,01 | 79,1 | 84,93 | 83,0 | 87,40 | 85,7 | 88,77 | 87,1 | 88,25 | 85,8 | 89,57 | 87,2 |

## CONCLUSÕES E TRABALHO FUTURO

O principal objetivo desta pesquisa foi desenvolver um método

para minerar dados educacionais, a fim de aprender um classificador para prever o sucesso de um estudo do aluno e verificar o método

em dados reais.

Empregamos métodos DM e SNA para resolver a tarefa. Nós verificou o método em alunos da Faculdade de Informática,

Universidade de Masaryk, mas os dados usados eram independente do corpo docente dente. Portanto, o método pode ser usado para qualquer unidade de um

universidade.

Mostramos que os dados estruturados obtidos por meio de

a análise de dados baseada em link aumentou a precisão da classe

## REFERÊNCIAS

* 1. R. Baker e K. Yacef. O estado dos dados educacionais mineração em 2009: uma revisão e visões futuras. Diário of Educational Data Mining, 1 (1): 3-17, 2009.
  2. J. Bayer, H. Bydzovská, J. Géryk, T. Obšıvac, e

L. Popelınský. Melhorar a classificação do estudo-

dados relacionados por meio de análise de rede social. Em pro- eventos da 7ª Oficina Doutoral em Matemática

e Métodos de Engenharia em Ciência da Computação, páginas 3-10. Universidade de Tecnologia de Brno, 2011.

* 1. J. Bayer, H. Bydzovská, J. Géryk e L. Popelınský.

sificação significativamente.

Usamos apenas os dados que não são específicos para um corpo docente. No entanto, para aumentar a precisão da classificação,

seria útil para enriquecer os dados com professores específicos em-

tributos, por exemplo, informações sobre exames específicos que um aluno foi aprovado ou reprovado. Outra forma possível de futuro melhoria pode ser explorar mais informações do

rede social.

Na verdade, usamos apenas informações sobre um aluno e seu ou seus vizinhos diretos. Foi intencional porque isso

a formação é fácil de obter e também fácil de incorporar em

o sistema de informação que é o objetivo desta pesquisa. Sobre por outro lado, relações mais complexas podem ajudar ainda mais aumentar o desempenho do sistema. Dados sobre comunicação

entre alunos e professores também pode ser útil. Lá-

portanto, planejamos construir uma rede heterogênea [6] onde o vértices serão de mais tipos. Diferentes métodos de aprendizagem pode ser usado então, por exemplo, classificação multi-rótulo.

## 6. AGRADECIMENTOS

Agradecemos a Michal Brandejs e todos os colegas do IS MU devel- equipe de apoio pelo suporte. Este trabalho foi parcialmente

apoiado pela Faculdade de Informática da Universidade de Masaryk.

Excalibur - uma ferramenta para mineração de dados. Em Processos de a Conferência Anual de Banco de Dados - Datakon 2011, páginas 227–228. Universidade de Tecnologia de Brno, 2011.

* 1. P. Carrington, J. Scott e S. Wasserman. Modelos e métodos de análise de redes sociais. Análise estrutural sis nas ciências sociais. Cambridge University Press, 2005.
  2. GW Dekker, M. Pechenizkiy e JM Vleeshouwers.

Previsão de abandono de alunos: um estudo de caso. Em EDM 2009: Anais da 2ª Conferência Internacional

Sobre mineração de dados educacional. Córdoba, Espanha., Páginas 41–50, 2009.

* 1. J. Han. Mineração de redes heterogêneas de informação explorando o poder dos links. Em Proceedings of the

20ª conferência internacional sobre aprendizagem algorítmica teoria, ALT'09, páginas 3-3, Berlin, Heidelberg, 2009.

Springer-Verlag.

* 1. J. Han, M. Kamber e J. Pei. Mineração de dados: conceitos e técnicas. A série Morgan Kaufmann em dados

Sistemas de gestão. Elsevier Science, 2011.

* 1. T. Kamada e S. Kawai. Um algoritmo para desenho

gráficos não direcionados gerais. Inf. Processo. Lett., 31: 7-15, Abril de 1989.

Anais da 5ª Conferência Internacional de Mineração de Dados Educacionais 108

# Página 7

* 1. S. Kotsiantis. Mineração de dados educacionais: um estudo de caso para prever alunos com tendência ao abandono. Int. J. Knowl.

Eng. Soft Data Paradigm., 1: 101-111, janeiro de 2009.

* 1. A. Kruger, A. Merceron e B. Wolf. Um modelo de dados

para facilitar a análise e mineração de dados educacionais. No EDM2010: Procedimentos da 3ª Conferência Internacional referência em Mineração de Dados Educacionais. Pittsburgh, EUA., páginas 131-140. [www.educationaldatamining.org,](http://www.educationaldatamining.org/) 2010.

* 1. F. Lemmerich, M. Ifland e F. Puppe. Identificando em

fatores de fluência no sucesso dos alunos por descoberta de subgrupo ery. Em EDM2011: Procedimentos da 4ª Internacional

Conferência sobre Mineração de Dados Educacionais. Eindhoven,

Holanda., páginas 345-346, 2011.

* 1. C. Marquez-Vera, C. Romero e S. Ventura. Pré-

eliminando o fracasso escolar usando mineração de dados. Em M. Pech- enizkiy, T. Calders, C. Conati, S. Ventura, C. Romero,

e JC Stamper, editores, EDM2011: Proceedings of

a 4ª Conferência Internacional de Dados Educacionais Mineração. Eindhoven, Holanda., Páginas 271–276. [www.educationaldatamining.org,](http://www.educationaldatamining.org/) 2011.

* 1. W. Nooy, A. Mrvar e V. Batagelj. Exploratório Social Análise de rede com Pajek. Análise Estrutural no

Ciências Sociais. Cambridge University Press, 2011.

* 1. C. Romero e S. Ventura. Mineração de dados educacionais: uma revisão do estado da arte. Trans. Sys. Man Cyber Parte C, 40: 601–618, novembro de 2010.
  2. A. Sahay e K. Mehta. Assistindo ensino superior

na avaliação, previsão e gerenciamento de questões relacionadas para o sucesso do aluno: um software baseado na web que usa dados mineração e implantação da função de qualidade. Acadêmico e

Conferência do Business Research Institute, 2010.

* 1. M. Cuhel, M. Brandejs, J. Kasprzak e T. Obšıvac.

Direitos de acesso na pesquisa de texto completo empresarial. No ICEIS 2010: Procedimentos da 12ª Conferência Internacional

ence on Enterprise Information Systems, Volume 1:

Integração de Bancos de Dados e Sistemas de Informação, páginas 32–39. INSTICC, Funchal, Portugal, 2010.

* 1. Y. Zhang, S. Oussena, T. Clark e H. Kim. Use dados

mineração para melhorar a retenção de alunos no ensino superior ção - um estudo de caso. No ICEIS 2010: Procedimentos do

12ª Conferência Internacional sobre Informação Empresarial Sistemas de ção, Volume 1: Bancos de dados e informações Integração de sistemas, páginas 190–197. INSTICC, Fun-

chal, Portugal, 2010.

Anais da 5ª Conferência Internacional de Mineração de Dados Educacionais 109