

Análise da Correlação da Evasão de Cursos de Graduação com o Empréstimo de Livros em Biblioteca

José G. de Oliveira Júnior, Robinson Vida Noronha, Celso Antônio Alves Kaestner

Universidade Tecnológica Federal do Paraná (UTFPR)
CEP 80.230-901 – Curitiba – PR – Brasil.

josjun@alunos.utfpr.edu.br, {vida, celsokaestner}@utfpr.edu.br

Abstract. *Data mining in high education is a recent field of research that is gaining popularity because of its potential for educational institutions, being a challenge to the scientific community on Computers in Education. The motivating factor of the search is to verify the correlation between the loan of books in the library with the dropout course. Therefore, this paper presents a model of knowledge extraction using time series and data mining. The experiments conducted with 3,605 undergraduate students at the Federal University of Technology - Paraná, said there is evidence of correlation between the loan of books in the library with the student will remain in the course.*

Resumo. *A mineração de dados no ensino superior é um campo de pesquisa recente que está ganhando popularidade por causa de seus potenciais para as instituições de ensino, sendo um desafio para comunidade científica de Informática na Educação. O fator motivador da pesquisa é verificar a existência de correlação entre o empréstimo de livros em biblioteca com o abandono de curso. Para isso, este artigo apresenta um modelo de extração de conhecimento utilizando séries temporais e mineração de dados. Os experimentos realizados com 3.605 alunos de graduação da Universidade Tecnológica Federal do Paraná indicam haver indícios da correlação entre o empréstimo de livros em biblioteca com a permanência do aluno no curso.*

1. Introdução e Motivação

Nos últimos anos têm surgido um crescente interesse e preocupação em identificar as causas do insucesso escolar e a identificação de padrões de comportamento dos alunos.

Uma alternativa de busca de informação muito promissora para atingir esse objetivo é o uso de descoberta de conhecimento em bases de dados e uso de técnicas de mineração de dados na educação, chamada Mineração de Dados Educacionais (*Educational Data Mining - EDM*) [Marquez-Vera et al., 2013]. Coordenadores de curso e gestores educacionais poderiam utilizar esses tipos de informação para definir estratégias que reduzissem os números de casos de insucesso escolar.

EDM é um campo que explora a estatística, a aprendizagem de máquina (*Machine Learning - ML*) e os algoritmos de mineração de dados (*Data Mining - DM*) sobre os diferentes tipos de dados da área de ensino. Seu principal objetivo é analisar estes tipos de dados, a fim de resolver questões em pesquisa educacional [Sachin e

Vijay, 2012]. EDM tem sido aplicada a grandes áreas de pesquisa, incluindo *e-learning*, sistemas tutores inteligentes, mineração de textos, mineração de redes sociais, etc. [Salzberg, 1997].

Com base no questionamento se existe alguma correlação entre o empréstimo de livros em biblioteca com o abandono de curso, este artigo descreve os resultados obtidos no uso de técnicas de DM aplicados em duas bases de dados. A primeira é de um Sistema de Controle Acadêmico e a segunda de um Sistema de Controle de Bibliotecas.

Para investigarmos esta possível correlação este artigo apresenta um modelo de extração do conhecimento em bases de dados educacionais, utilizando séries temporais, no levantamento dos atributos, e a aplicação de algoritmos de classificação em mineração de dados, para a identificação de padrões.

O restante do artigo está organizado da seguinte forma: na seção 2 estão elencados alguns trabalhos correlacionados com esta pesquisa; na seção 3 estão descritos os algoritmos de classificação utilizados nos experimentos; na seção 4 descreve-se o processo de seleção e tratamento dos dados; na seção 5 estão descritos os experimentos realizados e a análise dos resultados; e a seção 6 descreve as conclusões e possíveis trabalhos futuros.

2. Trabalhos Correlatos

Dekker et al. (2009) descreveram os resultados de um estudo de caso de mineração de dados educacional com a finalidade de prever a evasão de estudantes de Engenharia Elétrica, da Universidade de Eindhoven, após o primeiro semestre de seus estudos. Os resultados experimentais mostraram que classificadores bastantes simples e intuitivos (e. g. árvores de decisão) dão um resultado útil, com acurácia¹ entre 75 e 80%.

Lemos dos Santos et al. (2012) apresentam um estudo de caso sobre a utilização de técnicas de avaliações formativas a partir de um Ambiente Virtual de Aprendizagem. Resultados preliminares mostraram que a abordagem sugerida constitui-se em uma alternativa viável para um correto diagnóstico da evolução dos alunos e identificação dos alunos com maior nível de dificuldade na disciplina. A partir de dados coletados do ambiente foram criados modelos preditivos que permitem a identificação da tendência à reprovação com acurácia superior a 72%.

O trabalho de Paiva et al. (2013) trata da conceitualização e criação de uma ferramenta para a recomendação pedagógica cujo objetivo é prover aos professores de cursos baseados na web, recomendações pedagógicas personalizadas geradas por especialistas no domínio com base nos resultados da mineração dos dados educacionais dos alunos. Os resultados permitiram detectar padrões de interação úteis, usados na criação de recomendações, avaliadas por especialistas no domínio educacional/pedagógico, e disponibilizadas para que os professores as oferecessem a seus alunos.

O trabalho de Gottardo et al. (2012) aborda técnicas de mineração de dados educacionais utilizadas para geração de inferências sobre o desempenho de estudantes a

¹ Acurácia é uma medida de avaliação do desempenho de um modelo de classificação, que mede a taxa de acerto global, ou seja, o número de classificações corretas dividido pelo número total de instâncias dos dados a serem classificados.

partir de dados coletados em séries temporais. O objetivo principal foi investigar a viabilidade da obtenção destas informações em etapas iniciais de realização do curso, de forma a apoiar a tomada de ações pró-ativas. Os resultados obtidos demonstraram que é possível obter estas inferências com acurácia próxima a 75%, mesmo em períodos iniciais do curso.

3. Mineração de Dados Educacionais

Com o aumento do poder computacional e a ampliação do acesso virtual aos dados, estes são atualmente acumulados em um ritmo dramático. Há uma necessidade urgente de uma nova geração de teorias e ferramentas para ajudar os seres humanos na extração de informações úteis (conhecimento) dos volumes de rápido crescimento de dados digitais computacionais.

Essas teorias e ferramentas são o assunto do campo emergente da descoberta de conhecimento em bases de dados (*knowledge discovery in databases* - KDD). KDD é processo não trivial de identificação de padrões em dados que sejam válidos, inéditos, potencialmente úteis e compreensíveis. DM é a etapa no processo KDD que consiste na aplicação de algoritmos de descoberta que, considerando limitações computacionais aceitáveis, produzem uma enumeração particular de padrões (ou modelos) a partir de um conjunto de dados [Fayyad e Piatetsky-Shapiro, 1996].

A técnica de Mineração de Dados utilizada nesta pesquisa é conhecida como classificação. A partir de um conjunto de atributos previsores constrói-se um modelo que permite prever a classe de novas instâncias de dados. Os algoritmos de classificação possuem inúmeras aplicações, incluindo a detecção de fraudes, marketing de destino, previsão de desempenho, produção e diagnóstico médico [Han e Kamber, 2012].

Os classificadores investigados nesta pesquisa são: árvore de decisão, vizinho mais próximo, redes bayesianas, redes neurais e máquina de vetores de suporte.

Árvore de decisão é um conjunto de condições, organizados em uma estrutura hierárquica. É um modelo preditivo em que um exemplo é classificado, seguindo o caminho de condições satisfeitas, a partir da raiz da árvore até atingir uma folha, que vai corresponder a um rótulo de classe [Romero et al., 2008]. As árvores de decisão têm muitas vantagens: elas são simples e fáceis de entender, podem lidar com variáveis mistas (ou seja, variáveis tanto numéricas ou categóricas) [Hämäläinen e Vinni, 2011]. Quando uma árvore de decisão é construída, muitos dos ramos refletem anomalias nos dados de treinamento devido ao ruído ou desvios. Métodos de poda de árvores resolvem este problema de sobreajuste dos dados [Han e Kamber, 2012]. As árvores de decisão são consideradas modelos de fácil compreensão, porque um processo de raciocínio pode ser dado para cada conclusão, exceto se a árvore obtida é muito grande (uma série de nós e folhas) [Romero et al., 2008].

O classificador Vizinho Mais Próximo é uma técnica baseada em aprendizagem por analogia, ou seja, comparando uma determinada tupla teste com tuplas de treinamento que são semelhantes. As tuplas de treinamento são descritas por n atributos. Cada tupla representa um ponto em um espaço n -dimensional. Desta forma, todas as tuplas de formação são armazenadas num espaço padrão de n dimensões. Quando uma dada tupla é desconhecida, um classificador k -vizinho mais próximo procura o espaço

padrão para as tuplas de treinamento k que estão mais próximas da tupla desconhecida. Estas tuplas de treinamento k são os k “vizinhos mais próximos” da tupla desconhecida [Han e Kamber, 2012].

As redes bayesianas consideram a ideia que devemos levar em conta todas as informações sobre independências condicionais e representar uma estrutura de dependência mínima de atributos. Cada vértice do grafo corresponde a um atributo e as arestas definem o conjunto de atributos, da qual ele depende. A força de dependências é definida por probabilidades condicionais. Na tecnologia educacional, tem sido bastante comum para definir uma estrutura de grafo *ad hoc* por especialistas. No entanto, existe um risco elevado de que a rede resultante imponha dependências irrelevantes, pulando dependências efetivamente fortes [Hämäläinen e Vinni, 2011].

As Redes Neurais formam um paradigma também conhecido como redes de processamento paralelo distribuído. Trata-se de elementos de processamento interconectados chamados de nós ou neurônios que trabalham em conjunto para produzir uma função de saída. Redes Neurais *Feed-Forward* (RNFF) é o tipo mais utilizado de redes neurais, que possui a seguinte arquitetura em camadas de nós: uma para os nós de entrada, uma para nós de saída, e pelo menos uma camada de nós escondidos. Em cada camada oculta os nós são conectados aos nós da camada anterior e da camada seguinte, e as arestas estão associadas com pesos individuais. O modelo mais geral contém apenas uma camada oculta. RNFFs também podem representar qualquer tipo (não linear) de limites de classe [Hämäläinen e Vinni, 2011]. A principal desvantagem é que as RNFFs precisam de uma grande quantidade de dados, muito mais do que conjuntos típicos de dados educacionais contém. Elas são muito sensíveis ao sobreajuste, e o problema é ainda mais crítico com conjuntos de treinamento pequenos. O modelo de rede neural é do tipo “*black box*” (caixa preta) e é difícil para as pessoas entenderem as explicações para os resultados [Hämäläinen e Vinni, 2011].

O classificador “Máquina de Vetores de Suporte” (*Support Vector Machines - SVM*) é um método definido inicialmente para dados com separação linear. O objetivo é encontrar o hiperplano de maior margem que separa as classes. No caso de dados que não sejam linearmente separáveis utiliza-se o “Truque do *Kernel*” (*Kernel Trick*). No entanto, as SVMs têm a mesma restrição das redes neurais: os dados devem ser numéricos contínuos (ou quantificados); o modelo não é facilmente interpretável, e a seleção dos parâmetros adequados (especialmente a função de *kernel*) pode ser difícil [Hämäläinen e Vinni, 2011].

Para a extração de atributos para a mineração de dados foi utilizada uma série temporal, com algumas informações do vínculo do aluno com o curso. Série temporal é uma coleção de observações feitas sequencialmente ao longo do tempo. A característica mais importante deste tipo de dado é que as observações vizinhas são dependentes e estamos interessados em analisar e modelar esta dependência [Ehlers, 2007].

Nos experimentos realizados nesta pesquisa foram utilizados os seguintes algoritmos de classificação: J48 (árvore de decisão), Naive Bayes (redes bayesianas), SVM, usando a implementação SMO - *Sequential Minimal Optimization*, Multilayer Perceptron (redes neurais) e IBk (vizinho mais próximo).

4. Seleção e Tratamento dos Dados

Para a realização desta pesquisa, foram utilizadas as bases de dados de dois sistemas informatizados: o Sistema Acadêmico da UTFPR e o Sistema Integrado de Bibliotecas Pergamum.

O Sistema Acadêmico da UTFPR, em uso desde 1999, agrega as principais funções de controle e gestão acadêmica, abrangendo os portais (web) dos alunos e dos servidores.

O sistema integrado de bibliotecas Pergamum é utilizado pela UTFPR desde 2002. O sistema contempla as principais funções de uma Biblioteca, com o objetivo de facilitar a gestão dos centros de informação. O Pergamum é utilizado em mais de 400 instituições, aproximadamente 8.000 bibliotecas².

Para a obtenção dos atributos necessários à mineração, foi criada uma série temporal com algumas informações relevantes do vínculo do aluno com o curso, no final de cada ano e semestre. A série temporal resultou em um conjunto de 507.045 tuplas, entre os anos de 1979 a 2013, de 80.378 alunos dos cursos de graduação e técnico da instituição. Após a criação da série temporal, os dados para esta pesquisa foram selecionados, conforme a Tabela 2, com o seguinte escopo:

- Alunos de cursos de graduação, com oferta semestral, que ingressaram na UTFPR no primeiro semestre de 2012;
- Por ser um estudo preliminar, foram selecionados apenas os dados de quatro semestres (de 2012/1 a 2013/2).
- Com este escopo foram selecionados 3.605 alunos. A Tabela 1 exibe como estes alunos ficaram distribuídos em cada ano e semestre.

Tabela 1 – Quantitativo de alunos no final de cada ano e semestre

Ano	Semestre	Nº de alunos ativos	Nº de alunos desistentes	Total de alunos
2012	1º	2807	798	3605
2012	2º	2484	323	2807
2013	1º	2405	79	2484
2013	2º	2272	133	2405

As informações selecionadas do Sistema Acadêmico da UTFPR e do Sistema Pergamum estão apresentadas na Tabela 2.

² Fonte: http://www.pergamum.pucpr.br/redepergamum/pergamum_index.php

Tabela 2 – Atributos selecionados para a mineração de dados

Base de dados	Atributo	Descrição	Faixa de valores
Sistema Acadêmico da UTFPR	tipo_curso	Tipo de curso do aluno	Engenharia, Bacharelado, Tecnologia ou Licenciatura.
	genero	Gênero do aluno	Masculino ou feminino
	idade	Idade do aluno (no final de cada semestre letivo)	[17 .. 57]
	coeficiente	Coeficiente de rendimento do aluno: média de todas as disciplinas cursadas, ponderada pela carga horária total.	[0 .. 1]
	nota_linguagem	Nota do ENEM ³ (2011) em Linguagens, códigos e suas tecnologias	[327 .. 795]
	nota_humanas	Nota do ENEM (2011) em Matemática e suas tecnologias.	[289 .. 784]
	nota_natureza	Nota do ENEM (2011) em Ciências da Natureza e suas tecnologias;	[276 .. 844]
	nota_matematica	Nota do ENEM (2011) em Ciências Humanas e suas tecnologias	[321 .. 953]
	nota_redacao	Nota do ENEM (2011) em redação	[220 .. 1000]
	situacao_aluno	Situação do aluno no final do semestre. Ativos compreendem os alunos regulares, trancados, sub judice e em mobilidade acadêmica. Em desistentes estão incluídos os alunos que mudaram de curso.	Ativo ou Desistente
Pergamum	emp_biblioteca	Média de empréstimos de livros na biblioteca (por semestre)	[0 .. 113]

5. Experimentos Realizados

Para a realização desta pesquisa foi construído um modelo conceitual de extração do conhecimento utilizando séries temporais e mineração de dados, baseado no modelo de predição do insucesso escolar de Marquez-Vera et al. (2013). O modelo sugerido nesta pesquisa possui cinco etapas, conforme a Figura 1: (1) Coleta de dados: extração dos atributos das séries temporais para a utilização nos algoritmos de classificação; (2) Preparação dos dados: discretização dos atributos contínuos que forem necessários para a classificação; (3) Pré-processamento: seleção dos atributos utilizados nos algoritmos de mineração; (4) Mineração de dados: aplicação dos algoritmos de classificação; e (5) Interpretação: extração de conhecimento do modelo.

³ Exame Nacional do Ensino Médio, utilizado como forma de ingresso na UTFPR

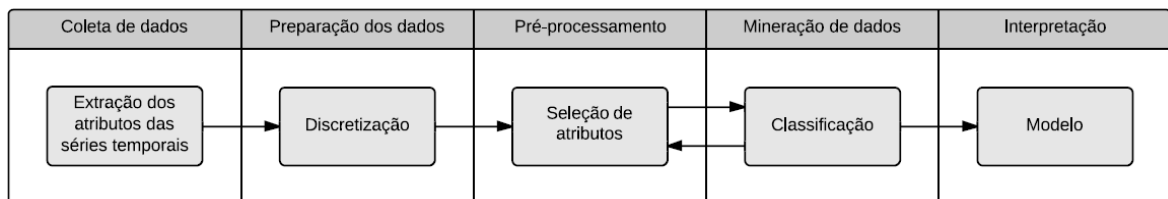


Figura 1 – Modelo de extração do conhecimento utilizando séries temporais e mineração de dados

O ambiente de mineração de dados WEKA foi utilizado nesta pesquisa, pois é reconhecido como um sistema de referência em mineração de dados e aprendizado de máquina [Hall et al., 2009]. Para a seleção dos conjuntos de treinamento e teste foi utilizada a técnica Validação Cruzada (*Cross Validation*). A ideia da Validação Cruzada é quebrar um conjunto de dados em k subconjuntos disjuntos de tamanhos aproximadamente iguais. Em seguida, executa-se k experimentos, onde, para cada experimento, o subconjunto de ordem k é removido. O sistema é treinado com o restante dos dados, e em seguida, o sistema de formação é testado no subconjunto mantido fora. No final desta validação cruzada k vezes, cada exemplo foi utilizado num teste definido uma única vez. Este procedimento tem a vantagem de que todos os conjuntos de teste são independentes [Witten et al., 2011].

Foram realizados dois experimentos: um experimento usando o atributo do coeficiente de rendimento do aluno e o segundo não utilizando este atributo.

5.1 Primeiro Experimento

O primeiro experimento utiliza o atributo coeficiente de rendimento do aluno. Para isso, este atributo foi discretizado, conforme a Tabela 3. As classes utilizadas levam em conta a correspondência usualmente classificadas pelos docentes que relacionam notas e conceitos.

Tabela 3 – Discretização do atributo coeficiente de rendimento do aluno

Título da Classe	Intervalo do coeficiente de rendimento	Número de alunos
A	[0.8 .. 1.0]	411
B	[0.7 .. 0.8[678
C	[0.6 .. 0.7[685
D	[0.5 .. 0.6[414
E	[0.2 .. 0.5[634
F	[0.0 .. 0.2[783

Analisando a árvore de decisão gerada pelo algoritmo J48, descrita na tabela 4, podemos verificar neste experimento que o atributo coeficiente do aluno é mais significativo que o número de empréstimos na biblioteca, pois para os alunos com idade maior que 18 anos e com coeficiente maior ou igual a 0.5 (classes A até D), independente de emprestar livros em biblioteca, o aluno é classificado na situação “Ativo”. Desta forma não é possível estabelecer uma correlação do empréstimo de livros em biblioteca com a desistência do curso. Ressalta-se que dos alunos maiores de 18

anos, com coeficiente menor que 0.5 (classes E e F), que não emprestaram livros, 88,8% desistiram do curso. Esses alunos representam 13,8% da amostra.

Tabela 4 – Árvore de decisão, gerada pelo algoritmo J48, do experimento 1

Regras	Valor da predição	Total de instâncias da regra	% da amostra	Instâncias classificadas incorretamente	Acurácia da regra
idade <= 18	Desistente	351	9,7%	26	92,6%
idade > 18					
emp_biblioteca = 0					
coeficiente = A	Ativo	34	0,9%	4	88,2%
coeficiente = B	Ativo	63	1,7%	18	71,4%
coeficiente = C	Ativo	65	1,8%	26	60,0%
coeficiente = D	Ativo	55	1,5%	17	69,1%
coeficiente = E	Desistente	137	3,8%	44	67,9%
coeficiente = F	Desistente	362	10,0%	12	96,7%
emp_biblioteca > 0					
coeficiente = A	Ativo	362	10,0%	29	92,0%
coeficiente = B	Ativo	607	16,8%	52	91,4%
coeficiente = C	Ativo	599	16,6%	42	93,0%
coeficiente = D	Ativo	347	9,6%	35	89,9%
coeficiente = E					
genero = M	Ativo	257	7,1%	75	70,8%
genero = F					
emp_biblioteca <= 17					
nota_matematica <= 527	Ativo	30	0,8%	7	76,7%
nota_matematica > 527					
nota_linguagem <= 595	Desistente	49	1,4%	13	73,5%
nota_linguagem > 595	Ativo	67	1,9%	29	56,7%
emp_biblioteca > 17	Ativo	22	0,6%	4	81,8%
coeficiente = F	Desistente	198	5,5%	15	92,4%

5.2 Segundo Experimento

O segundo experimento não utiliza o atributo coeficiente de rendimento do aluno. Neste experimento pode-se analisar, conforme a Tabela 5, que 80,6% dos alunos com mais de 18 anos que emprestaram livros em biblioteca não haviam desistido do curso, durante os quatro semestres analisados. Esses alunos representam 70,4% da amostra. Esta informação é um indício que o empréstimo de livros pode estar relacionado com a permanência do aluno no curso.

Tabela 5 - Árvore de decisão, gerada pelo algoritmo J48, do experimento 2

Regras	Valor da predição	Total de instâncias da regra	% da amostra	Instâncias classificadas incorretamente	Acurácia da regra
idade <= 18	Desistente	351	9,7%	26	92,6%
idade > 18					
emp_biblioteca = 0					
tipo_curso = Bacharelado					
nota_matematica <= 780	Desistente	92	2,6%	28	69,6%
nota_matematica > 780	Ativo	27	0,7%	7	74,1%
tipo_curso = Engenharia	Desistente	209	5,8%	74	64,6%
tipo_curso = Tecnologia	Desistente	247	6,9%	70	71,7%
tipo_curso = Licenciatura	Desistente	141	3,9%	16	88,7%
emp_biblioteca > 0	Ativo	2538	70,4%	492	80,6%

5.3 Comparação dos experimentos

Nos experimentos cada algoritmo foi executado 10 vezes, usando a validação cruzada e considerou-se a média aritmética dessas execuções como o desempenho final do algoritmo, indicado à Tabela 6.

Tabela 6 – Acurácia média e seu desvio padrão obtidos nos experimentos

Classificador	Experimento 1	Experimento 2
J48	86.61% ± 1.60	79.84% ± 1.98
Multilayer Perceptron	84.26% ± 1.78	75.39% ± 2.79
SMO	83.10% ± 1.67	66.07% ± 2.35
IBk	82.88% ± 1.90	63.23% ± 2.33
Naive Bayes	80.07% ± 1.88	65.49% ± 2.22

De acordo com os resultados obtidos, o algoritmo de árvore de decisão J48 apresentou a melhor acurácia e menor desvio padrão, conforme mostrado na tabela 6. Isto é oportuno, pois a árvore de decisão permite que se visualizem os relacionamentos entre os atributos previsores que levaram a determinada classificação.

Nos dois experimentos foi utilizado como critério de poda, no algoritmo J48, somente os nós que possuíam a partir de 20 instâncias, para evitar que a árvore ficasse muito grande, dificultando o entendimento do modelo (parâmetro *minNumObj*).

Ambos os experimentos permitiram que se descobrisse uma informação não investigada: 92,6% dos alunos com idade entre 17 e 18 anos desistiram do curso.

6. Conclusão e Trabalhos Futuros

Esta pesquisa apresentou uma abordagem de extração do conhecimento combinando séries temporais e mineração de dados. A criação das séries temporais permitiu a seleção dos atributos na janela de tempo definida para este estudo, facilitando o trabalho de levantamento e agregação dos dados.

Foram aplicados cinco algoritmos de classificação e o melhor desempenho foi obtido com o classificador de árvore de decisão J48, com acurácia próxima a 80%. A informação que 80,6% dos alunos, com mais de 18 anos, que emprestaram livros na biblioteca estão ativos no curso é um indício da correlação entre estes dois atributos. É importante destacar que em sua essência os algoritmos de mineração de dados encontram correlações entre os dados considerados, e não relações do tipo causa-efeito. No caso, os resultados indicam que existe uma correlação entre a desistência de um aluno de seu curso e o não empréstimo de livros. As causas destes fatos, no entanto, podem ser diversas, e estão fora do escopo deste trabalho. A hipótese proposta foi comprovada: se um aluno realiza poucos empréstimos de livros, há um forte indicativo de que o mesmo desista de seu curso no futuro.

Estes resultados podem indicar a importância do acesso à biblioteca para a continuidade dos alunos em um curso, o que remete a uma reflexão sobre a importância da disponibilidade de materiais didáticos adicionais para os cursos não presenciais.

Assim, espera-se ter contribuído com a comunidade de Informática na Educação pela construção de um modelo conceitual que possibilita a extração do conhecimento em bases de dados educacionais usando séries temporais e mineração de dados. Como trabalhos futuros pretende-se tratar o problema dos dados desbalanceados, o uso de um

conjunto maior de atributos combinado com algoritmos de seleção de atributos e a aplicação de classificação sensível ao custo.

Referências

- Dekker, G. W., Pechenizkiy, M., Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. International Working Group on Educational Data Mining.
- Ehlers, R. S. (2007). Análise de séries temporais. Laboratório de Estatística e Geoinformação. Universidade Federal do Paraná. Disponível em <http://www.icmc.usp.br/pessoas/ehlers/stemp/stemp.pdf>
- Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37.
- Gottardo, E., Kaestner, C., Noronha, R. V. (2012). Previsão de Desempenho de Estudantes em Cursos EAD Utilizando Mineração de Dados: uma Estratégia Baseada em Séries Temporais. In *Anais do Simpósio Brasileiro de Informática na Educação* (Vol. 23, No. 1).
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1), 10-18.
- Hämäläinen, W., Vinni, M. (2011). Classifiers for educational data mining. *Handbook of Educational Data Mining*, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 57-71.
- Han, J.; Kamber, M.(2012). *Data mining concepts and techniques*, third edition. Morgan Kaufmann Publishers.
- Lemos dos Santos, H., Nunes Prates Camargo, F., Silva Camargo, S. (2012). Predizendo o sucesso de estudantes através do uso avaliações formativas em AVAs. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação* (Vol. 1, No. 1).
- Marquez-Vera, C., Morales, C. R., & Soto, S. V. (2013). Predicting School Failure and Dropout by Using Data Mining Techniques. *Tecnologias del Aprendizaje*, IEEE Revista Iberoamericana de, 8(1), 7-14.
- Paiva, R., Bittencourt, I. I., da Silva, A. P. (2013). Uma Ferramenta para Recomendação Pedagógica Baseada em Mineração de Dados Educacionais. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação* (Vol. 1, No. 1).
- Romero, C., Ventura, S., Espejo, P. G., Hervás, C. (2008, June). Data Mining Algorithms to Classify Students. In *EDM* (pp. 8-17).
- Sachin, R. B., Vijay, M. S. (2012, January). A survey and future vision of data mining in educational field. In *Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on* (pp. 96-100). IEEE.
- Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data mining and knowledge discovery*, 1(3), 317-328.
- Witten, I. H., Frank, E., (2011). *Data Mining: Practical machine learning tools and techniques*, third edition. Morgan Kaufmann Publishers.