Criação e Seleção de Atributos Aplicados na Previsão da Evasão de Curso em Alunos de Graduação

José G. de Oliveira Júnior[1], Robinson Vida Noronha[1], Celso A. Alves Kaestner[1]

[1]Universidade Tecnológica Federal do Paraná (UTFPR)
CEP 80.230-901 - Curitiba - PR - Brasil

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

Abstract. *One of the challenges of educational institutions is to reduce the course dropout. A very promising solution to achieve this goal is the use of educational data mining in order to identify patterns that assist managers in decision making. This paper proposes a dropout prediction model using the creation and selection of attributes derived from educational databases. The experiments were applied to undergraduate students of a public institution of higher education. The experimental results show the most relevant attributes to predict the dropout, indicating the contribution of the feature creation in the data mining task.*

Resumo. *Um dos desafios das instituições de ensino é reduzir o abandono de curso. Uma solução muito promissora para atingir esse objetivo é o uso da mineração de dados educacionais, a fim de identificar padrões que auxiliem os gestores na tomada de decisão. Este trabalho propõe um modelo de previsão da evasão escolar utilizando a criação e seleção de atributos oriundos de bases de dados educacionais. Os experimentos foram aplicados em alunos de graduação de uma Instituição Pública de Ensino Superior. Os resultados experimentais apresentam os atributos mais relevantes para prever a evasão, indicando a contribuição da criação de atributos na tarefa de mineração de dados.*

1. Introdução

Detectar antecipadamente quais estudantes não terão êxito na conclusão do curso tem sido um grande desafio para a comunidade acadêmica e pesquisadores da área de educação. Essa possível detecção antecipada poderia fornecer informações que permitissem a tomada de decisões de gestores acadêmicos (por exemplo: coordenadores de curso, diretores de ensino, entre outros) para modificar essa predição detectada. Na busca por um dispositivo ou mecanismo inteligente que seja capaz de realizar essa detecção antecipada, alguns pesquisadores da área de Informática em Educação têm empregado técnicas computacionais de mineração de dados. Nesse contexto, bases de dados acadêmicas (por exemplo: Sistema de Controle Acadêmico e Ambientes Virtuais de Aprendizagem) têm sido investigadas por meio de algoritmos de mineração de dados [Baker et al. 2011], [Gottardo et al. 2014], [Rigo et al. 2012] e [Borges et al. 2015].

No contexto de mineração de dados, a criação de atributos consiste em criar novos atributos a partir de outros existentes, de modo que informações importantes sejam capturadas em um conjunto de dados mais eficazmente.

A seleção de atributos é uma técnica aplicada para reduzir a dimensionalidade dos dados, facilitando a aplicação de algoritmos de mineração. A redução de dimensionalidade produz uma representação mais compacta, mais facilmente interpretável conceito alvo, focalizando a atenção do usuário sobre as variáveis mais relevantes [Witten et al. 2011]. O problema da seleção de atributos pode ser definido como encontrar um subconjunto de atributos de um conjunto de dados original que produza um classificador com melhor acurácia.

Neste trabalho é proposto um modelo de previsão da evasão escolar, utilizando classificação, criação e seleção de atributos, com a finalidade de auxiliar a análise da evasão de alunos de cursos presenciais de graduação.

O restante do artigo está organizado da seguinte forma: na seção 2 são apresentados os trabalhos relacionados a esta pesquisa; na seção 3 é apresentado o modelo proposto para a previsão da evasão; na seção 4 estão descritos os experimentos realizados; e finalmente a seção 5 apresenta as conclusões e trabalhos e os futuros.

2. Trabalhos Relacionados

[Kotsiantis et al. 2003] realizaram uma série de experimentos com dados fornecidos pelos cursos de informática da Hellenic Open University, com o objetivo de identificar o algoritmo de aprendizado mais adequado para efetuar a previsão do abandono de curso. A comparação de seis algoritmos de classificação mostrou que o algoritmo Naïve Bayes foi o mais adequado. Os resultados obtiveram acurácia de 63%, baseado somente em dados demográficos, e acurácia de 83% antes da metade do período letivo.

[Dekker et al. 2009] apresentam os resultados de um estudo de caso de mineração de dados educacionais com a finalidade de prever a evasão de estudantes do curso 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 acurácia entre 75 e 80%.

[Manhães et al. 2011] comparam seis algoritmos de classificação e apresentam uma abordagem quantitativa, aplicados em uma base de dados de informações acadêmicas de alunos de graduação da UFRJ (Universidade Federal do Rio de Janeiro), com o objetivo de identificar precocemente alunos em risco de evasão. Os melhores resultados foram obtidos com o algoritmo Naïve Bayes, obtendo acurácia em torno de 80%.

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 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, para apoiar a tomada de ações. Os resultados obtidos demonstraram que é possível obter inferências com acurácia próxima a 75%, utilizando os algoritmos "RandomForest" e "MultilayerPerceptron", com dados originários apenas dos períodos iniciais do curso.

3. Modelo de previsão da evasão

O modelo proposto neste trabalho, conforme mostrado na Figura 1, é baseado nos trabalhos de [Fayyad et al. 1996], [Márquez-Vera et al. 2013] e [Chau e Phung 2013]. O modelo é composto de 7 etapas que são indicadas a seguir.

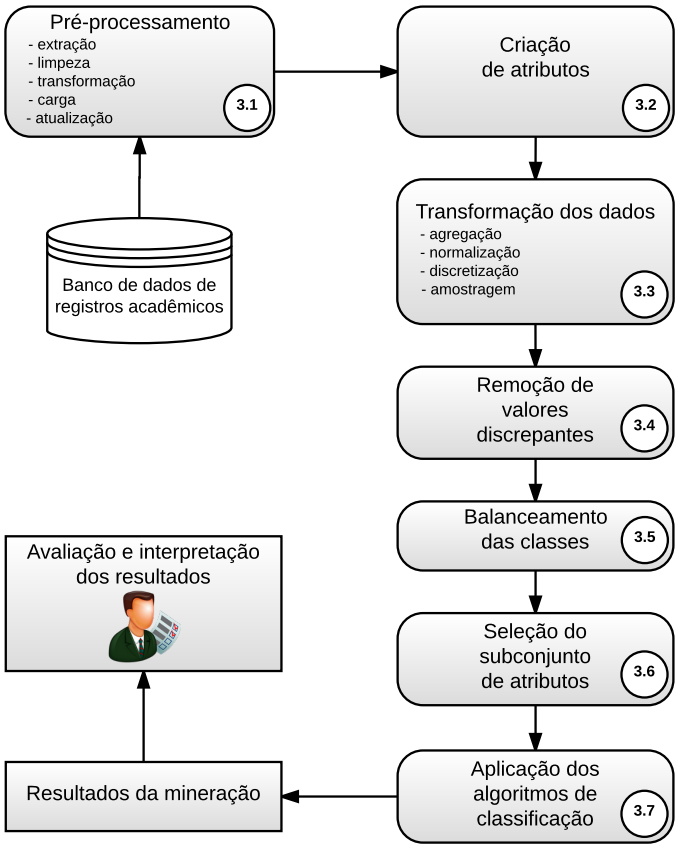

Figura 1. Modelo proposto de previsão da evasão escolar

3.1. Pré-processamento

Nesta etapa são realizadas as atividades de extração dos dados, limpeza, transformação, carga e atualização dos dados, conforme os procedimentos tradicionais empregados em mineração de dados [Fayyad et al. 1996].

3.2. Criação de Atributos

A criação de novos atributos pode capturar informações importantes em um conjunto de forma mais eficiente do que os atributos originais. Este trabalho propõe a criação de novos atributos, detalhados na Seção 4.1, considerando informações existentes na base de dados e utilizando medidas estatísticas para a sua definição. O objetivo dos novos atributos é criar índices quantitativos que sejam simples e fáceis de calcular, e que sirvam como "sinais de alerta" para os gestores educacionais, permitindo a tomada de ações a tempo de evitar a evasão.

3.3. Transformação dos Dados

Nesta etapa são realizadas as tarefas de agregação, normalização, discretização e amostragem dos dados, também seguindo os procedimentos tradicionais empregados em mineração de dados, como descrito em [Han et al. 2011].

3.4. Remoção dos Valores Discrepantes

Nesta etapa é verificada a necessidade de remoção de valores discrepantes (*outliers*). Um forma de localizarmos os valores discrepantes pode ser feita com a aplicação do cálculo da amplitude interquartil, definida pela diferença entre o $1^{o}(Q1)$ e o $3^{o}(Q3)$ quartil. Os

limites superiores e inferiores são calculados conforme expressão apresentada abaixo, e os valores fora destes limites são considerados valores discrepantes e eliminados:

$$LimiteInferior = max\{min(dados); Q_1 - outlier\_factor \times (Q_3 - Q_1)\} \quad (1)$$

$$LimiteSuperior = min\{max(dados); Q_3 + outlier\_factor \times (Q_3 - Q_1)\} \quad (2)$$

3.5. Balanceamento das Classes

Apesar da evasão ser um problema nas instituições de ensino, o número de casos de evasão ainda é pequeno em relação ao número total de alunos. Sendo assim o problema se caracteriza pelo desbalanceamento das classes. Este problema faz com que os algoritmos de aprendizagem tendam a ignorar as classes menos frequentes (classes minoritárias) e só considerar nas mais frequentes (classes majoritárias). Como resultado, o classificador não é capaz de classificar corretamente as instâncias de dados correspondentes a classes pouco representadas [Márquez-Vera et al. 2013].

Uma abordagem amplamente utilizada no balanceamento de classes é a aplicação do algoritmo SMOTE (*Synthetic Minority Oversampling Technique*) [Chawla et al. 2002]. Esse algoritmo, empregado neste trabalho, ajusta a frequência relativa entre classes majoritárias e minoritárias, introduzindo sinteticamente instâncias de classes minoritárias, considerando a técnica K-nn [Witten et al. 2011].

3.6. Seleção do Subconjunto de Atributos

O problema da seleção de um subconjunto de atributos (*feature subset selection*) é encontrar um subconjunto de atributos originais de um conjunto de dados, de tal forma que um algoritmo de indução, que é executado nos dados contendo apenas esses atributos, gere um classificador com a maior acurácia possível. A seleção do subconjunto de atributos possui duas abordagens principais: *filter* e *wrapper*.

A abordagem *filter* seleciona os atributos usando uma etapa de pré-processamento. A principal desvantagem dessa abordagem é que ela ignora totalmente os efeitos do subconjunto de atributos selecionados no desempenho do algoritmo de indução [Kohavi e John 1997].

Na abordagem *wrapper*, proposta por [John et al. 1994], o algoritmo de seleção do subconjunto de atributos existe como um invólucro em torno do algoritmo de indução. A ideia por trás da abordagem *wrapper* é simples. O algoritmo de indução é executado no conjunto de dados, geralmente dividido em conjuntos de treinamento e validação. O subconjunto de atributos com a maior acurácia é escolhido como o último conjunto no qual se deve executar o algoritmo de indução. O classificador resultante é, então, avaliado em um conjunto de teste independente que não foi usado durante a pesquisa.

3.7. Aplicação dos Algoritmos de Classificação

Neste trabalho foi empregado o método de classificação, que é o processo de colocação de um objeto específico (conceito) em um conjunto de categorias, com base nas respectivas propriedades do objeto. Para a mineração de dados educacionais é recomendado o uso de algoritmos do tipo "caixa branca", que geram modelos de fácil interpretação e podem ser usados diretamente para a tomada de decisão [Márquez-Vera et al. 2013]. Os principais classificadores nesta categoria são os baseados em regras (JRip), árvore de decisão (J48) e modelagem estatística (Naïve Bayes), todos disponíveis na ferramenta WEKA.

4. Experimentos

Os experimentos foram realizados com dados extraídos do sistema acadêmico da UTFPR, sendo selecionados os dados de alunos ingressantes pelo SISU dos cursos presenciais de graduação com oferta semestral, conforme atributos descritos na Tabela 1.

Foi utilizado nos experimentos o ambiente de mineração de dados WEKA, reconhecido como um sistema de referência em mineração de dados e aprendizado de máquina [Hall et al. 2009].

4.1. Criação de Atributos

Foram criados 11 atributos, conforme indicado à Tabela 1, que estão detalhados a seguir. Para os experimentos foram criados 3 *datasets*, em que foram empregados dados de alunos coletados durante 6 semestres letivos.

Tabela 1. Atributos utilizados nos experimentos

Nº	Atributo	Tipo	Atributo criado
01	grau (engenharia, bacharelado, tecnologia ou licenciatura)	Categórico	
02	genero (masculino ou feminino)	Categórico	
03	estado_civil	Categórico	
04	tipo_escola_anterior (pública ou privada)	Categórico	
05	reentrada_mesmo_curso (sim/não)	Categórico	Sim
06	mudou_de_curso (sim/não)	Categórico	Sim
07	tipo_cota	Categórico	
08	previsao_evasao_dificuldade_disciplinas_cursadas (sim/não)	Categórico	Sim
09	idade_inicio_curso	Numérico	
10	total_semestres_trancados	Numérico	Sim
11	emprestimos_biblioteca_por_semestre	Numérico	Sim
12	regressao_coeficiente	Numérico	Sim
13	percentual_frequencia	Numérico	Sim
14	coeficiente_rendimento	Numérico	
15	percentual_aprov	Numérico	Sim
16	nota_final_enem	Numérico	
17	nota_linguagem	Numérico	
18	nota_humanas	Numérico	
19	nota_natureza	Numérico	
20	nota_matematica	Numérico	
21	nota_redacao	Numérico	
22	micro_regiao_origem (mesma do câmpus ou outra)	Categórico	Sim
23	meso_regiao_origem (mesma do câmpus ou outra)	Categórico	Sim
24	regiao_origem (mesma do câmpus ou outra)	Categórico	Sim
25	socio_renda_familiar	Categórico	
26	socio_mora_com	Categórico	
27	socio_reside_em	Categórico	
28	socio_trabalho	Categórico	
29	socio_necessidade_trabalhar	Categórico	
30	socio_part_economica_na_familia	Categórico	
31	socio_escolaridade_pai	Categórico	
32	socio_escolaridade_mae	Categórico	
33	socio_tipo_escola	Categórico	
34	socio_fez_cursinho	Categórico	
35	socio_motivo_escolha_curso	Categórico	
36	evasao (sim/não) [atributo alvo]	Categórico	

4.1.1. Dificuldade Média das Disciplinas Cursadas pelo Aluno

O primeiro atributo proposto utiliza o conceito de dificuldade de uma disciplina/turma cursada pelos alunos, definido pela relação inversa do percentual de aprovação dos alunos na disciplina/turma, definido por:

$$Dif(d) = \log_2 \left(\frac{Ap(d) + Rep(d)}{Ap(d)} \right) \tag{3}$$

em que:

$Dif(d) \rightarrow$ dificuldade de aprovação na disciplina/turma (d) em um período letivo;
$Ap(d) \rightarrow$ número de alunos aprovados na disciplina/turma;
$Rep(d) \rightarrow$ número de alunos reprovados na disciplina/turma;

A partir daí é possível computar o atributo denominado de "dificuldade média das disciplinas cursadas pelo aluno". Esse atributo agrega um componente coletivo (percentual dos alunos aprovados na disciplina/turma) ao desempenho individual do aluno. O cálculo do valor deste atributo é feito com a seguinte equação:

$$DM(a) = \frac{\sum_{i=1}^{n} Dif(D_n) - \sum_{j=1}^{m} Dif(D_m)}{n + m} \tag{4}$$

em que:

$DM(a) \rightarrow$ dificuldade média das disciplinas cursadas pelo aluno (a);
$n \rightarrow$ total de disciplinas que o aluno obteve aprovação;
$m \rightarrow$ total de disciplinas que o aluno reprovou;
$D_n \rightarrow$ disciplina que o aluno obteve aprovação;
$D_m \rightarrow$ disciplina que o aluno reprovou.

Na investigação de quais seriam os valores aceitáveis de dificuldade média das disciplinas cursadas pelos alunos, foi aplicado o cálculo em uma amostra composta de 16.766 alunos de cursos semestrais de graduação formados na UTFPR entre os anos de 1983 e 2014, que ingressaram no curso no 1º período. Foram utilizados os alunos formados por serem a referência de desempenho acadêmico de sucesso.

Conforme observado por [Mendes Braga et al. 2003], a evasão é mais intensa nos períodos iniciais dos cursos. Sendo assim, procurou-se identificar em que período do curso a evasão acumulada atingisse 80%, para concentrar a análise da evasão nos períodos mais críticos. Foi investigada esta informação nos cursos semestrais com 6, 8 ou 10 semestres. Verificou-se que aproximadamente 80% das desistências acontecem até o 3º período do curso, independente do total de períodos do curso. Sendo assim, foi selecionado na amostra somente as disciplinas cursadas até o 3º período do curso. Segue abaixo o resumo dos dados estatísticos do atributo de dificuldade média de disciplinas cursadas pelos alunos formados:

Mínimo	-0,7500	Máximo	1,5500	Mediana	0,2200
1º Quartil	0,0700	3º Quartil	0,3600	Média	0,2220

Para a amostra selecionada o valor da variância interquartil foi de 0,28. Desta forma, o intervalo para exclusão dos valores discrepantes, utilizando $1,5 \times IQR$, são os valores situados fora do intervalo [-0.37 .. 0.80], resultando em 2,73% de valores discrepantes na amostra selecionada. Ou seja, os alunos que neste atributo estiverem fora desse intervalo poderão ser considerados "em risco de evasão".

4.1.2. Demais Atributos Criados

- Reentrada no mesmo curso: indica se o aluno está reiniciando o mesmo curso, no mesmo câmpus.
- Mudou de curso: indica se o aluno é oriundo de outro curso de graduação da instituição.
- Total de semestres trancados: indica a quantidade de semestres em que o aluno esteve com a matrícula trancada.
- Empréstimos na biblioteca: indica a média de empréstimos de livros na biblioteca por semestre cursado.
- Regressão do coeficiente de rendimento: indica o coeficiente angular (β) da equação de regressão linear do coeficiente de rendimento médio das disciplinas cursadas em cada semestre.
- Percentual de frequência: indica o percentual de frequência das disciplinas cursadas.
- Percentual de Aprovação: indica o percentual de aprovação das disciplinas cursadas.
- Micro, meso e região de origem dos calouros: estes três atributos indicam se o aluno é oriundo da mesma microrregião, mesorregião ou região (IBGE) do câmpus.

4.2. Normalização e Remoção dos Valores Discrepantes

Nos dados utilizados foram normalizados os valores de notas do SISU para o intervalo [0.00,1.00]. Para a criação dos *datasets* foram removidos os valores discrepantes do atributo idade, utilizando a variação interquartil com $outlier\_factor = 3$.

4.3. Balanceamento das Classes

Para o balanceamento de classes foi aplicado o algoritmo SMOTE [Chawla et al. 2002], com os percentuais de instâncias sintéticas inseridas conforme indicado na Tabela 2.

Tabela 2. Distribuição de classes do atributo alvo

Dataset	Selecionados	Não Evadidos	Evadidos	% de evadidos	% de instâncias inseridas
DS1	2436	1498	938	38,51%	59,70%
DS2	2694	1626	1068	39,64%	52,24%
DS3	3325	1885	1440	43,31%	30,90%

4.4. Seleção do Subconjunto de Atributos

Os algoritmos de seleção de atributos utilizados nos experimentos são os disponíveis na ferramenta WEKA [Witten et al. 2011]:
– Abordagem *filter*: CfsSubsetEval, ChiSquaredAttributeEval, GainRatioAttributeEval, InfoGainAttributeEval, OneRAttributeEval, ReliefFAttributeEval e SymmetricalUncertAttributeEval;
– Abordagem *wrapper*: WrapperSubsetEval, utilizando o classificador de árvore de decisão (J48).

Depois de selecionados os subconjuntos de atributos, o classificador J48 foi executado usando a validação cruzada (fator $n = 10$). Para esta etapa foi utilizado o ambiente

Weka Experiment Environment [Hall et al. 2009], utilizando o meta classificador FilteredClassifier. Com esse meta classificador foi aplicado o filtro "attribute.Remove" (para a seleção do subconjunto de atributos) e posteriormente o classificador J48. Para os subconjuntos selecionados pelos algoritmos de filtro foram selecionados apenas os 10 melhores atributos ranqueados.

A acurácia e o seu desvio padrão da aplicação do classificador J48 nos subconjuntos selecionados estão mostrados na Tabela 3.

Tabela 3. Acurácia e seu desvio padrão obtidos com o classificador J48 nos subconjuntos de atributos

Algoritmo de seleção	Método de busca	DS1	DS2	DS3
ChiSquaredAttributeEval	Ranking	83,59 ± 2,04	83,72 ± 2,02	85,71 ± 1,97 *
GainRatioAttributeEval	Ranking	83,67 ± 2,11	83,65 ± 1,98	85,64 ± 2,03 *
InfoGainAttributeEval	Ranking	83,59 ± 2,04	83,72 ± 2,02	85,71 ± 1,97 *
SymmetricalUncert	Ranking	83,59 ± 2,04	83,72 ± 2,02	85,75 ± 2,03 *
OneRAttributeEval	Ranking	83,74 ± 1,94	83,24 ± 2,07	85,71 ± 1,97 *
ReliefFAttributeEval	Ranking	83,17 ± 1,91	81,99 ± 2,05 *	84,52 ± 2,00 *
WrapperSubsetEval	BestFirst	83,94 ± 2,02	83,81 ± 2,04 **	87,15 ± 1,73
WrapperSubsetEval	GeneticSearch	84,60 ± 2,59 **	81,36 ± 2,02 *	87,26 ± 1,79 **
CfsSubsetEval	BestFirst	83,86 ± 2,08	81,82 ± 2,09 *	83,94 ± 2,02 *
CfsSubsetEval	GeneticSearch	83,65 ± 2,10	83,75 ± 1,98	83,88 ± 2,01 *

Após a obtenção da acurácia do classificador J48 para cada um dos subconjuntos de atributos, foram desprezados os atributos selecionados em que a acurácia não obteve significância estatística, quando comparados ao melhor resultado obtido.

Para se obter os melhores atributos foi utilizado o seguinte procedimento: 1) ordenou-se de forma decrescente a frequência em que o atributo foi selecionado pelos algoritmos WrapperSubsetEval e CfsSubsetEval; 2) ordenou-se de forma crescente pela posição média que o atributo foi ordenado pelos algoritmos que utilizam o ranqueamento. O resultado dessa seleção está indicado na Tabela 4.

Tabela 4. Classificação dos melhores atributos na mineração

Classificação	Nº do atributo	Atributo	Novo atributo	Nº de vezes selecionado**	Posição média*
1º	12	regressao_coeficiente	Sim	8	8
2º	8	previsao_evasao_dificuldade_disciplinas_cursadas	Sim	7	2
3º	14	coeficiente_rendimento		6	4
4º	15	percentual_aprov	Sim	5	3
5º	10	total_semestres_trancados	Sim	5	20
6º	11	emprestimos_biblioteca_por_semestre	Sim	3	8
7º	27	socio_reside_em		3	13
8º	1	grau		3	15
9º	32	socio_escolaridade_mae		2	7
10º	29	socio_necessidade_trabalhar		2	11

* posição média nos algoritmos de seleção de atributos por ranqueamento

** frequência que o atributo foi selecionado nos algoritmos Wrapper e Cfs

4.5. Análise dos resultados

Preliminarmente à aplicação dos algoritmos de mineração, foi verificado que aproximadamente 80% da evasão de curso ocorre até o 3º período, independente se o curso possui duração de 6, 8 ou 10 períodos.

Nos três *datasets* utilizados a abordagem wrapper obteve a melhor acurácia, com resultados entre 83 e 87%.

Com os resultados apresentados na Tabela 4 pode-se concluir que a criação de atributos contribuiu para a tarefa de mineração de dados. Dos dez melhores atributos classificados, cinco deles são novos atributos. Estes novos atributos podem facilitar a tarefa de análise da evasão com o objetivo reduzi-la. O atributo de dificuldade média das disciplinas cursadas pelo aluno revelou-se uma boa medida de prognóstico de desempenho do aluno, tendo um componente coletivo em sua avaliação.

5. Conclusão e Trabalhos Futuros

Esta pesquisa apresentou um modelo de previsão da evasão escolar, utilizando a criação de novos atributos e a seleção dos melhores atributos previsores. O algoritmo de seleção de atributos que apresentou os melhores resultados para a acurácia foi o WrapperSubsetEval, que utiliza a abordagem *wrapper*, empregando o classificador de árvore de decisão J48. Este resultado é consistente com o indicado em [Hall e Holmes 2003], em que a abordagem wrapper também aparece com os melhores resultados.

Dos seis melhores atributos selecionados para a tarefa de mineração, cinco deles foram novos atributos, indicando a sua contribuição na tarefa de previsão da evasão. A criação do atributo "dificuldade média das disciplinas cursadas pelo aluno" melhorou a acurácia dos algoritmos de classificação, agregando um componente coletivo (percentual dos alunos aprovados na disciplina) no desempenho individual do aluno.

Com o modelo de previsão da evasão proposto, espera-se proporcionar aos gestores educacionais indicadores e/ou um conjunto de regras que permitam avaliar a possibilidade da evasão de cada aluno. Como trabalhos futuros pretende-se aplicar o modelo em outras amostras e também avaliar a aplicação da classificação sensível ao custo.

Referências

Baker, R., Isotani, S., e Carvalho, A. (2011). Mineraçao de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, 19(02):03.

Borges, V. A., Nogueira, B. M., e Barbosa, E. F. (2015). Uma análise exploratória de tópicos de pesquisa emergentes em informática na educação. *Revista Brasileira de Informática na Educação*, 23(01):85.

Chau, V. T. N. e Phung, N. H. (2013). Imbalanced educational data classification: An effective approach with resampling and random forest. In *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2013 IEEE RIVF International Conference on*, pages 135–140. IEEE.

Chawla, N. V., Bowyer, K. W., Hall, L. O., e Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Dekker, G. W., Pechenizkiy, M., e Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. *International Working Group on Educational Data Mining*.

Fayyad, U., Piatetsky-Shapiro, G., e Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37.

Gottardo, E., Kaestner, C., e 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*, volume 23.

Gottardo, E., Kaestner, C. A. A., e Noronha, R. V. (2014). Estimativa de desempenho acadêmico de estudantes: Análise da aplicação de técnicas de mineração de dados em cursos a distância. *Revista Brasileira de Informática na Educação*, 22(01):45.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., e Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hall, M. A. e Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *Knowledge and Data Engineering, IEEE Transactions on*, 15(6):1437–1447.

Han, J., Kamber, M., e Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

John, G. H., Kohavi, R., Pfleger, K., et al. (1994). Irrelevant features and the subset selection problem. In *ICML*, volume 94, pages 121–129.

Kohavi, R. e John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324.

Kotsiantis, S. B., Pierrakeas, C., e Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 267–274. Springer.

Liu, H. e Motoda, H. (1998). *Feature extraction, construction and selection: A data mining perspective*. Springer Science & Business Media.

Manhães, L. M. B., Cruz, S. d., Costa, R. J. M., Zavaleta, J., e Zimbrão, G. (2011). Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. *Anais do Simpósio Brasileiro de Informática na Educação*.

Márquez-Vera, C., Morales, C. R., e Soto, S. V. (2013). Predicting school failure and dropout by using data mining techniques. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 8(1):7–14.

Mendes Braga, M., Peixoto, M. D. C. L., e Bogutchi, T. F. (2003). A evasão no ensino superior brasileiro: O caso da ufmg. *Avaliação*, 8(3):161–189.

Rigo, S. J., Cazella, S. C., e Cambruzzi, W. (2012). Minerando dados educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. In *Anais do Workshop de Desafios da Computação Aplicada à Educação*, pages 168–177.

Witten, I. H., Frank, E., e Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.