Revista de Informática Aplicada, Volume 13, Número 2, 2017

Método de Seleção de Atributos Aplicados na Previsão da Evasão de Cursos de Graduação

José G. de Oliveira Júnior [1], Robinson Vida Noronha [2], Celso A. Alves Kaestner [3],

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 educacionais na tomada de decisão. Este trabalho detalha um método de seleção de atributos aplicados na previsão da evasão escolar utilizando criação e seleção de atributos oriundos de bases de dados educacionais. Os experimentos foram realizados com dados de alunos de graduação extraídos do sistema acadêmico da Universidade Tecnológica Federal do Paraná. 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. A abordagem é genérica e pode ser aplicada a uma grande quantidade de instituições de ensino.

Palavras-chave: Mineração de Dados Educacionais. Criação de Atributos. Seleção de Subconjunto de Atributos

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 educational managers in decision making. This paper details a method of selecting attributes applied in dropout prediction using the creation and selection of attributes from educational databases. The experiments were appliedt with data from undergraduate students extracted from the academic system of the Federal University of Technology - Paraná. The experimental results present the most relevant attributes to predict evasion, indicating the contribution of attribute creation in the data mining task. The approach is generic and can be applied to a large number of educational institutions.

Keywords: Feature Subset Selection, Students' Dropout, Educational Data Mining

[1]Universidade Tecnológica Federal do Paraná (UTFPR), Av Sete de Setembro, 3165 - Curitiba - PR, E-mail: josjun@alunos.utfpr.edu.br

[2]Universidade Tecnológica Federal do Paraná (UTFPR), Av Sete de Setembro, 3165 - Curitiba - PR, E-mail: vida@utfpr.edu.br

[3]Universidade Tecnológica Federal do Paraná (UTFPR), Av Sete de Setembro, 3165 - Curitiba - PR, E-mail: celsokaestner@utfpr.edu.br

1 Introdução

As instituições de ensino superior, tanto as públicas quanto as privadas, têm como desafio reduzir o abandono de curso, melhorando assim os indicadores de retenção e conclusão de curso. 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 por gestores acadêmicos (e.g. coordenadores de curso, diretores de ensino, entre outros) para modificar essa predição detectada. Na busca por um dispositivo ou mecanismo 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 têm sido investigadas por meio de algoritmos de mineração de dados (BAKER; ISOTANI; CARVALHO, 2011), (GOTTARDO; KAESTNER; NORONHA, 2014), (RIGO; CAZELLA; CAMBRUZZI, 2012) e (BORGES; NOGUEIRA; BARBOSA, 2015).

A previsão da evasão escolar pode ser associada à tarefa de mineração de dados chamada classificação, que tem como objetivo a associação de uma classe a cada elemento considerado, a partir de um conjunto de propriedades (ou atributos previsores) inerentes ao mesmo elemento.

Dentre as técnicas empregadas em mineração de dados destacam-se a criação e a seleção de atributos (KOHAVI; JOHN, 1997). 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 aplicada para reduzir a dimensionalidade dos dados, facilitando a aplicação de algoritmos de mineração. A redução de dimensionalidade produz uma representação mais compacta, focalizando a atenção do usuário sobre as variáveis mais relevantes (WITTEN; FRANK; HALL, 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[4].

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

Este trabalho é uma extensão do artigo "Criação e Seleção de Atributos Aplicados na Previsão da Evasão de Curso em Alunos de Graduação" (JÚNIOR; KAESTNER; NORONHA, 2016) apresentado no *Computer on the Beach* 2016.

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 método proposto para a seleção de atributos; na seção 4 estão descritos os experimentos realizados; e finalmente a seção 5 apresenta as conclusões e trabalhos e os futuros.

2 Trabalhos Relacionados

Pesquisadores da área de Informática na Educação têm aplicado conceitos de mineração de dados com o objetivo de identificar padrões na busca de modelos de previsão da evasão de curso.

É o caso, por exemplo, do trabalho de Kotsiantis, Pierrakeas e Pintelas (2003), onde foram realizados 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

[4] 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 a serem classificadas.

dados demográficos, e acurácia de 83% antes da metade do período letivo.

Com a finalidade de prever a evasão de estudantes do curso de Engenharia Elétrica, da Universidade de Eindhoven, Dekker, Pechenizkiy e Vleeshouwers (2009) apresentam os resultados de um estudo de caso de mineração de dados educacionais 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%.

Com o objetivo de identificar precocemente alunos em risco de evasão 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). Os melhores resultados foram obtidos com o algoritmo Naïve Bayes, obtendo acurácia em torno de 80%.

O trabalho de Gottardo, Kaestner e Noronha (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 Método para Seleção dos Melhores Atributos

Este trabalho propõe um método de seleção de atributos aplicados na previsão da evasão de alunos matriculados em cursos de graduação, conforme mostrado na Figura 1, que é baseado nos trabalhos de (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), (MÁRQUEZ-VERA; MORALES; SOTO, 2013) e (CHAU; PHUNG, 2013). O método é composto de 10 etapas que são detalhadas a seguir.

3.1 Pré-processamento

Na etapa 1, Figura 1, são realizadas as atividades de extração, limpeza, transformação, carga e atualização dos dados, conforme os procedimentos tradicionais empregados em mineração de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Nesta pesquisa os dados foram persistidos em um *Data Warehouse*, pois facilitaram bastante os trabalhos de mineração, principalmente nesta etapa de pré-processamento.

3.2 Criação de Atributos

A criação de novos atributos, etapa 2 da Figura 1, pode capturar informações importantes em um conjunto de dados de forma mais eficiente do que os atributos originais. Este trabalho propõe a criação de novos atributos, apresentados 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 decisões a tempo de evitar a evasão.

3.3 Transformação dos Dados

Na etapa 3, Figura 1, 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; KAMBER; PEI, 2011).

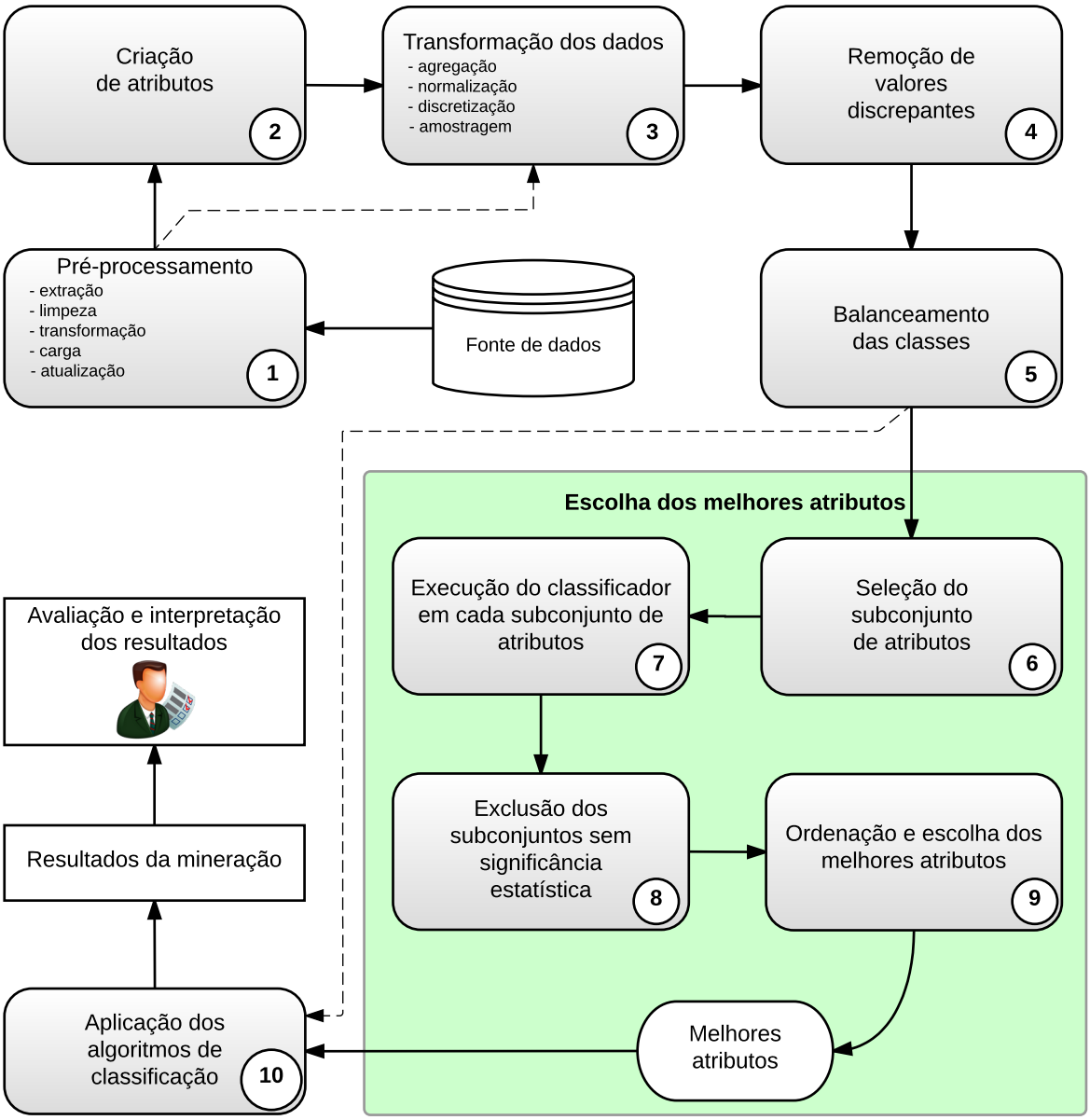

Figura 1: Método de seleção dos melhores atributos para classificação

3.4 Remoção dos Valores Discrepantes

Na etapa 4, Figura 1, é verificada a necessidade de remoção de valores discrepantes (*outliers*) que possam influenciar negativamente no resultado obtido com a mineração de dados. Um forma de se localizar os valores discrepantes é com a aplicação do cálculo amplitude interquartil (*Inter Quartile Range* - IQR), definida pela diferença entre o 1º($Q1$) e o 3º($Q3$) 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 menor em relação ao número total de alunos ativos no curso. 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 as 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; MORALES; SOTO, 2013).

Uma abordagem comumente utilizada no balanceamento de classes, etapa 5 da Figura 1, 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; FRANK; HALL, 2011).

3.6 Seleção do Subconjunto de Atributos

A seleção de atributos, etapa 6 da Figura 1, é um método de redução da dimensionalidade onde são detectados e removidos atributos irrelevantes, fracamente relevantes ou redundantes (HAN; KAMBER; PEI, 2011). Selecionar 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* (KOHAVI; JOHN, 1997).

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; 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 Execução do Classificador em Cada Subconjunto de Atributos

Depois de selecionados os subconjuntos de atributos, os classificadores devem ser avaliados quanto ao desempenho, utilizando-se como medida a acurácia. Nos experimentos realizados, os algoritmos foram executados dez vezes nos subconjuntos selecionados, usando a técnica de validação cruzada (fator $n = 10$).

Na etapa 7, Figura 1, foi utilizado nos experimentos o ambiente WEKA Experiment Environment - WEE (HALL et al., 2009), aplicando-se o meta classificador FilteredClassifier. Nesse meta classificador foi aplicado o filtro "attribute.Remove" (para a seleção do subconjunto de atributos) e posteriormente foi aplicado o classificador. Para os subconjuntos selecionados pelos algoritmos com a abordagem *filter* foram selecionados apenas os dez melhores atributos ranqueados.

3.8 Exclusão dos Subconjuntos sem Significância Estatística

O objetivo da etapa 8, Figura 1, é descartar o subconjunto de atributos cuja acurácia seja muito inferior à melhor acurácia obtida com o classificador no experimento.

Para avaliar a significância estatística dos resultados obtidos, utiliza-se a técnica de teste estatístico conhecida como "T-pareado" (*pairwise T-test*) (WITTEN; FRANK; HALL,

2011), com nível de significância de 5%.

A partir do resultado do teste "T-pareado", considerando o nível de significância de 5%, são desprezados os subconjuntos de atributos selecionados em que a acurácia não obteve significância estatística, quando comparados com a melhor acurácia obtida (denominada *test base* no WEE). Caso todos os subconjuntos selecionados pela abordagem *filter* não obtenham significância estatística, deve-se selecionar o subconjunto com a melhor acurácia, para assim permitir realizar o desempate na próxima etapa.

3.9 Ordenação e Escolha dos Melhores Atributos

Na etapa 9, Figura 1, para se obter os melhores atributos utilizou-se o seguinte procedimento:

1. Ordena-se de forma decrescente a frequência em que o atributo foi selecionado pelos algoritmos WrapperSubsetEval e CfsSubsetEval, que não utilizam o método de busca *Ranking*;

2. Ordena-se de forma crescente pela posição média em que o atributo foi classificado pelos algoritmos que utilizaram o método de busca Ranking.

3. Selecionam-se os *n* melhores atributos.

3.10 Aplicação dos Algoritmos de Classificação

A etapa 10, Figura 1, apresenta a aplicação dos algoritmos de classificação. Neste trabalho foi empregado o procedimento de classificação, que é a atribuição de um 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; MORALES; SOTO, 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 - *Waikato Environment for Knowledge Analysis* (HALL et al., 2009).

4 Experimentos

Os experimentos foram realizados com dados extraídos do sistema acadêmico da Universidade Tecnológica Federal do Paraná (UTFPR), sendo selecionados os dados de alunos ingressantes pelo SISU (Sistema de Seleção Unificada, gerenciado pelo Ministério da Educação) dos cursos presenciais de graduação com oferta semestral, conforme atributos descritos na Tabela 2.

Nos experimentos foi utilizado 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

Para os experimentos foram criados 3 *datasets*, em que foram empregados dados de alunos coletados durante 6 semestres letivos, conforme detalhado na Tabela 1.

Tabela 1: *Datasets* criados para os experimentos

Dataset	Ingressantes	Ano/Semestre fim	Total de alunos	*Outlier*	% de *outlier*
DS1	2010/1	2012/2	2566	130	5,07%
DS2	2011/1	2013/2	2847	153	5,37%
DS3	2012/1	2014/2	3438	113	3,29%

Foram criados 11 atributos, conforme indicado na Tabela 2, que estão detalhados a seguir.

Tabela 2: 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 atributo nº 8 da Tabela 2 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:

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

em que:

$Dif(d) \rightarrow$ dificuldade de aprovação na disciplina/turma (d);
$Ap(d) \rightarrow$ 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:

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

em que:

$DM(a) \rightarrow$ dificuldade média das disciplinas cursadas por um 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 (BRAGA; PEIXOTO; BOGUTCHI, 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. Os desistentes por período de cursos semestrais de graduação estão ilustrados nos gráficos das Figuras 2, 3 e 4. Analisando esses gráficos pode-se verificar 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".

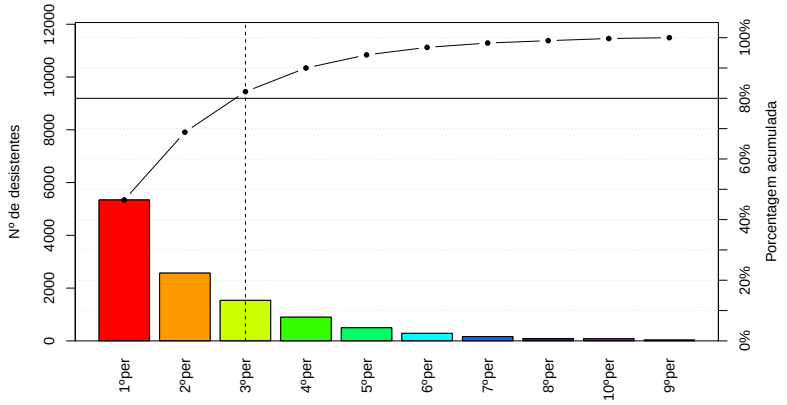

Figura 2: Desistentes por período em cursos de graduação com 6 Semestres (1981 a 2014)

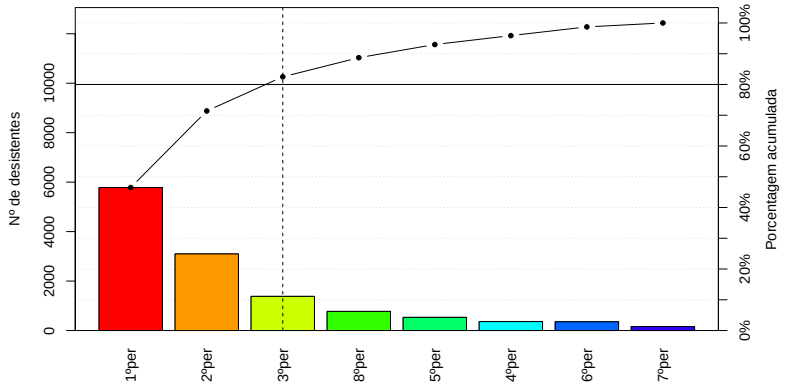

Figura 3: Desistentes por período em cursos de graduação com 8 semestres (1999 a 2014)

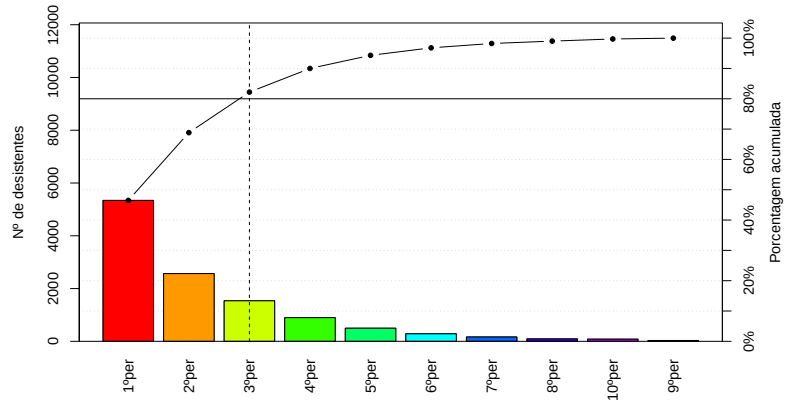

Figura 4: Desistentes por período em cursos de graduação com 10 semestres (1978 a 2014)

4.1.2 Demais Atributos Criados

- Atributo nº 5 da Tabela 2: Reentrada no mesmo curso. Indica se o aluno está reiniciando o mesmo curso, no mesmo câmpus.

- Atributo nº 6 da Tabela 2: Mudou de curso. Indica se o aluno é oriundo de outro curso de graduação da instituição.

- Atributo nº 10 da Tabela 2: Total de semestres trancados. Indica a quantidade de semestres em que o aluno esteve com a matrícula trancada.

- Atributo nº 11 da Tabela 2: Empréstimos na biblioteca. Indica a média de empréstimos de livros na biblioteca por semestre cursado.

- Atributo nº 12 da Tabela 2: 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. Esse atributo permite analisar a tendência de melhora, piora ou manutenção do coeficiente de rendimento do aluno.

- Atributo nº 13 da Tabela 2: Percentual de frequência. Indica o percentual de frequência das disciplinas cursadas.

- Atributo nº 15 da Tabela 2: Percentual de Aprovação. Indica o percentual de aprovação das disciplinas cursadas.

- Atributos nºs 22, 23 e 24 da Tabela 2: 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 3.

Tabela 3: Distribuição de classes do atributo alvo (atributo nº 36 da Tabela 2)

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; FRANK; HALL, 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 são mostrados na Tabela 4.

Tabela 4: 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 *

Tabela 5: 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

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 5.

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 5 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 de alertar os gestores para o problema e para a necessidade de buscar soluções pra reduzir a evasão. O atributo de dificuldade média das disciplinas cursadas pelo aluno revelou-se uma boa medida de prognóstico de desempenho do aluno, uma vez que possui um componente coletivo em sua avaliação.

5 Conclusão e Trabalhos Futuros

Esta pesquisa apresentou um método de seleção dos melhores atributos aplicados em algoritmos de classificação para a previsão da evasão em cursos de graduação, 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; 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.

Este trabalho apresenta uma extensão do artigo (JÚNIOR; KAESTNER; NORONHA, 2016), e buscou fazer um melhor detalhamento da etapa de seleção dos melhores atributos, subdividindo-a em 4 etapas: Seleção do Subconjunto de Atributos, Execução do Classificador em Cada Subconjunto de Atributos, Exclusão dos Subconjuntos sem Significância Estatística e Ordenação e Escolha dos Melhores Atributos.

Com o método 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 método em outras amostras e também avaliar a aplicação da classificação sensível ao custo.

Referências

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineraçao de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, v. 19, n. 02, p. 03, 2011.

BORGES, V. A.; NOGUEIRA, B. M.; BARBOSA, E. F. Uma análise exploratória de tópicos de pesquisa emergentes em informática na educação. *Revista Brasileira de Informática na Educação*, v. 23, n. 01, p. 85, 2015.

BRAGA, M. M.; PEIXOTO, M. D. C. L.; BOGUTCHI, T. F. A evasão no ensino superior brasileiro: O caso da ufmg. *Avaliação*, Unicamp, v. 8, n. 3, p. 161–189, 2003.

CHAU, V. T. N.; PHUNG, N. H. Imbalanced educational data classification: An effective approach with resampling and random forest. In: IEEE. *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2013 IEEE RIVF International Conference on*. [S.l.], 2013. p. 135–140.

CHAWLA, N. V. et al. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002.

DEKKER, G. W.; PECHENIZKIY, M.; VLEESHOUWERS, J. M. Predicting students drop out: A case study. *International Working Group on Educational Data Mining*, ERIC, 2009.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37, 1996.

GOTTARDO, E.; KAESTNER, C.; NORONHA, R. V. 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*. [S.l.: s.n.], 2012. v. 23, n. 1.

GOTTARDO, E.; KAESTNER, C. A. A.; NORONHA, R. V. 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*, v. 22, n. 01, p. 45, 2014.

HALL, M. et al. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009.

HALL, M. A.; HOLMES, G. Benchmarking attribute selection techniques for discrete class data mining. *Knowledge and Data Engineering, IEEE Transactions on*, IEEE, v. 15, n. 6, p. 1437–1447, 2003.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123814790, 9780123814791.

JOHN, G. H. et al. Irrelevant features and the subset selection problem. In: *ICML*. [S.l.: s.n.], 1994. v. 94, p. 121–129.

JÚNIOR, J. G. O.; KAESTNER, C.; NORONHA, R. V. Criação e seleção de atributos aplicados na previsão da evasão de curso em alunos de graduação. *Anais do Computer on the Beach*, p. 061–070, 2016.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. *Artificial intelligence*, Elsevier, v. 97, n. 1, p. 273–324, 1997.

KOTSIANTIS, S. B.; PIERRAKEAS, C.; PINTELAS, P. E. Preventing student dropout in distance learning using machine learning techniques. In: SPRINGER. *Knowledge-Based Intelligent Information and Engineering Systems*. [S.l.], 2003. p. 267–274.

MANHÃES, L. M. B. et al. Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. *Anais do Simpósio Brasileiro de Informática na Educação*, 2011.

MÁRQUEZ-VERA, C.; MORALES, C. R.; SOTO, S. V. Predicting school failure and dropout by using data mining techniques. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, IEEE, v. 8, n. 1, p. 7–14, 2013.

RIGO, S. J.; CAZELLA, S. C.; CAMBRUZZI, W. Minerando dados educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. In: *Anais do Workshop de Desafios da Computação Aplicada à Educação*. [S.l.: s.n.], 2012. p. 168–177.

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd. ed. San Francisco,

CA, USA: Morgan Kaufmann Publishers Inc.,
2011. ISBN 0123748569, 9780123748560.