Análise do Desempenho Acadêmico Utilizando Redes Bayesianas: um estudo de caso

Danilo Raniery Alves Coutinho, Thereza Padilha

Departamento de Ciências Exatas - Universidade Federal da Paraíba (UFPB)
Campus IV – Rio Tinto – PB – Brasil

{danilo.coutinho, thereza}@dce.ufpb.br

Abstract. *This paper shows a case study involving the analysis of students performance in the Computational Applied Logic course, available in the Computer Science Course, at UFPB-Campus IV, using the Bayesian networks technique. To obtain the construction of the Bayesian network and, hence, probabilistic inferences the GeNIe 2.1 tool was used. Results with Greedy Thick Thinning and Bayesian Search algorithms are showed.*

Resumo. *Este artigo apresenta um estudo de caso envolvendo a análise do desempenho dos alunos na disciplina de Lógica Aplicada à Computação do curso de Licenciatura em Ciência da Computação da UFPB-Campus IV, utilizando a técnica de Redes Bayesianas. Para realizar a construção da rede Bayesiana e, consequentemente, as inferências probabilísticas, utilizou-se a ferramenta GeNIe 2.1. Resultados utilizando os algoritmos Greedy Thick Thinning e Bayesian Search são apresentados.*

1. Introdução

Atualmente, há grande quantidade de dados armazenados por professores, de forma digital, nas suas disciplinas e, cada vez mais, ferramentas para análise destes dados estão sendo disponíveis. Comparar notas, frequências ou até fazer associações entre desempenhos acadêmicos de turmas não é uma tarefa muito fácil. Com o aumento de pesquisas envolvendo as áreas de Mineração de Dados (MD) e Educação, surgiu uma nova área denominada mineração de dados educacionais (*Educational Data Mining* - EDM) que visa encontrar informações úteis a partir de dados educacionais. A MD envolve a aplicação de algoritmos que exploram os dados, o desenvolvimento de modelos e o descobrimento de padrões ou informações úteis ocultas em grandes bases de dados [Marques 2014]. Nesta área, destaca-se a mineração de dados por meio de redes Bayesianas (RB), devido a possibilidade de tratar domínios que envolvem graus de incerteza. Assim, pode-se encontrar modelos que representem o domínio com relacionamentos de causa-efeito e, sobretudo, informações probabilísticas.

À vista disso, este trabalho tem como objetivo analisar e compreender probabilisticamente variáveis e/ou eventos que influenciaram o desempenho acadêmico dos alunos que cursaram a disciplina Lógica Aplicada à Computação entre os semestres 2013.2 a 2014.2, da UFPB-campus IV. Para isso, o foco está voltado para a tarefa de mineração de causas utilizando dois algoritmos para a geração automática de RB, disponíveis na ferramenta GeNIe, para encontrar informações ocultas sobre o

desempenho acadêmico e, com isso, auxiliar os professores da disciplina e a gestão acadêmica em ações pedagógicas.

Este artigo encontra-se organizado da seguinte forma: na seção 2 é abordado o processo de descoberta de conhecimento de dados, incluindo suas etapas. Na seção 3 é apresentada uma introdução sobre Redes Bayesianas. Na seção 4 são mostrados os resultados dos experimentos realizados, bem como a descrição do conjunto de dados utilizado. Por fim, na seção 5, são apresentadas as considerações finais.

2. Mineração de Dados Educacionais

MDE é uma área de pesquisa que tem por objetivo encontrar informações ocultas em bases de dados de domínio da Educação para, assim, auxiliar a tomada de decisões [Baker et al. 2011]. Como esta área surgiu a partir de estudos de mineração de dados, muitos trabalhos seguem o processo de descoberta de conhecimento de dados (*Knowledge Discovery in Database - KDD*), composto por 5 etapas, que são [Fayyad et al. 1996]:

- **seleção**: a primeira etapa do processo KDD e prevê a escolha do conjunto de dados a ser minerado, associado a um domínio, contendo todos os seus atributos (também chamados de características/variáveis) e registros (também chamados de exemplos). Estes dados podem estar disponíveis em fontes diferentes;
- **pré-processamento**: nesta etapa procura-se trabalhar o conjunto de dados a fim de eliminar dados redundantes e inconsistentes, recuperar dados incompletos e discretizar dados contínuos (reduzir quantidade de valores possíveis). Esta etapa é vista como a mais dispendiosa, mas o alvo é obter um conjunto de dados o mais consistente possível para a mineração;
- **transformação**: esta etapa visa formatar os dados pré-processados na etapa anterior para que sejam reconhecidos pelos algoritmos de mineração a serem aplicados;
- ***data mining* (mineração de dados)**: esta etapa prevê a exploração e análise dos algoritmos de mineração de dados utilizando os conjuntos de dados disponibilizados com objetivo de identificar padrões (conhecimento);
- **interpretação/avaliação**: esta etapa está voltada para a interpretação e avaliação dos padrões identificados na etapa anterior.

As tarefas de mineração de dados educacionais podem ser qualificadas da seguinte forma: predição (classificação, regressão ou estimação de densidade), agrupamento (*clustering*), mineração de relações (associação, correlação, sequência ou causal), modelagem de dependência, entre outras [Baker et al. 2011]. Cada tarefa está relacionada com o tipo de conhecimento que se deseja extrair, ou seja, o alvo da mineração de dados. A mineração de causas, que é o foco deste trabalho, busca verificar se um evento influencia a ocorrência de outro evento através de análise de padrões. Por exemplo, se um aluno apresenta performance ruim em uma prova isso pode decorrer por falta de domínio de conteúdos anteriores ou por dificuldade de aprendizagem. Assim, a mineração de causa tenta inferir eventos com variáveis buscando situações de causa-efeito.

3. Redes Bayesianas

Segundo [Costa 2013] apud [Coppin 2010], redes Bayesianas são grafos orientados acíclicos em que os nós representam evidências ou hipóteses (variáveis) e os arcos

representam a dependência probabilística direta entre os nós. Assim, uma RB consiste em uma parte qualitativa (estrutura causal entre as variáveis do domínio) e outra quantitativa (em que cada nó dispõe de uma tabela de probabilidade condicional em que representa as probabilidades de cada estado dado o estado dos seus nós-pais).

A construção das redes Bayesianas pode ocorrer de forma manual (a partir do conhecimento de especialistas do domínio) ou automática (através de conjuntos de dados utilizando algoritmos de aprendizado). Três ferramentas comuns que trabalham com a construção de redes Bayesianas automaticamente são: GeNIe [Bayesfusion 2016], Bayesware [Borges 2005] e RapidMiner [Rapidminer 2016]. Para a realização deste trabalho, foi utilizada a ferramenta GeNIe 2.1, da BayesFusion, que encontra-se disponível gratuitamente (versão acadêmica) no endereço http://www.bayesfusion.com. GeNIe reconhece formatos como .dat, .txt, .csv ou arquivo ODBC para a entrada dos dados e dispõe de algoritmos de aprendizado como *Bayesian Search*, *Greedy Thick Thinning* e *Naive Bayes*.

4. Experimentos Realizados

Nesta seção é apresentado e discutido a aplicação das cinco etapas do processo KDD para os dados dos alunos matriculados disciplina de Lógica Aplicada à Computação.

4.1. Seleção

O conjunto de dados utilizado nos experimentos refere-se ao desempenho dos alunos matriculados na disciplina de Lógica Aplicada à Computação, 64 horas, durante os semestres 2013.2 (25 alunos), 2014.1 (39 alunos) e 2014.2 (47 alunos), da Universidade Federal da Paraíba, campus IV. O professor da disciplina disponibilizou, para cada semestre, uma planilha eletrônica contendo a identificação do aluno, as notas obtidas nas avaliações realizadas durante a disciplina, a quantidade de faltas e a sua situação no final da disciplina.

4.2. Pré-processamento

A Tabela 1 apresenta os dez atributos que compõem o conjunto de dados selecionado, bem como seus valores possíveis e descrição. Todos os valores contínuos foram discretizados para que pudessem ser lidos pela ferramenta GeNIe 2.1. Por exemplo, no caso do atributo Frequência, que indica a quantidade de faltas dos alunos na disciplina, foram criados quatro intervalos de, aproximadamente, 15 horas (n0_15, n16_31, n32_47, n48_64) e, consequentemente, todos os seus valores foram substituídos. Os valores dos atributos Trabalho X e Prova X, que possuía valores entre 0 a 3,0 e 0 a 7,0, respectivamente, também tiveram tratamento similar.

Tabela 1. Atributos do conjunto de dados

Nome	Valores Possíveis	Descrição
Semestre	s20132, s20141 e s20142	Representa o semestre cursado pelo aluno
Frequência	n0_15, n16_31, n32_47 e n48_64	Representa a quantidade de faltas
Resultado	aprovado, reprovado e desistente	Representa a situação do aluno no final da disciplina
Trabalho1	n0_15 e n16_30	Representa a nota recebida pelo aluno referente aos trabalhos da unidade 1

Prova1	n0_20, n21_49 e n50_70	Representa a nota recebida pelo aluno referente à prova da unidade 1
Trabalho2	n0_15 e n16_30	Representa a nota recebida pelo aluno referente aos trabalhos da unidade 2
Prova 2	n0_20, n21_49 e n50_70	Representa a nota recebida pelo aluno referente à prova da unidade 2
Trabalho3	n0_15 e n16_30	Representa a nota recebida pelo aluno referente aos trabalhos da unidade 3
Prova 3	n0_20, n21_49 e n50_70	Representa a nota recebida pelo aluno referente à prova da unidade 3

Nas planilhas analisadas, não foram encontrados dados ausentes ou inconsistentes. Como a identificação de cada aluno não é relevante para a mineração de dados, esta foi ignorada.

4.3. Transformação

Para que o conjunto de dados fosse reconhecido pela ferramenta GeNIe 2.1, os dados foram formatados para um arquivo com extensão .txt, com variáveis e valores separados por tabulações.

4.4. *Data mining* (mineração de dados) e Interpretação/Avaliação

Foram aplicados vários algoritmos de aprendizado disponíveis, que fornecem rede Bayesiana como resultado, para o conjunto de dados carregado. Neste trabalho serão apresentados e discutidos resultados de experimentos utilizando os algoritmos *Greedy Thick Thinning* e *Bayesian Search*, pois foram os que tiveram melhor resposta.

4.4.1. Algoritmo *Greedy Thick Thinning*

A Figura 1 apresenta a estrutura construída na ferramenta usando o algoritmo de aprendizagem *Greedy Thick Thinning*. Observou-se que o algoritmo identificou implicações (arcos) do nó Prova 1 nos nós Prova 2 e Prova 3 o que, de fato, realmente ocorre no mundo real, uma vez que o conteúdo desta disciplina é acumulativo. O nó Trabalho 3 influenciou na nota da Prova 3 e o resultado final do aluno na disciplina.

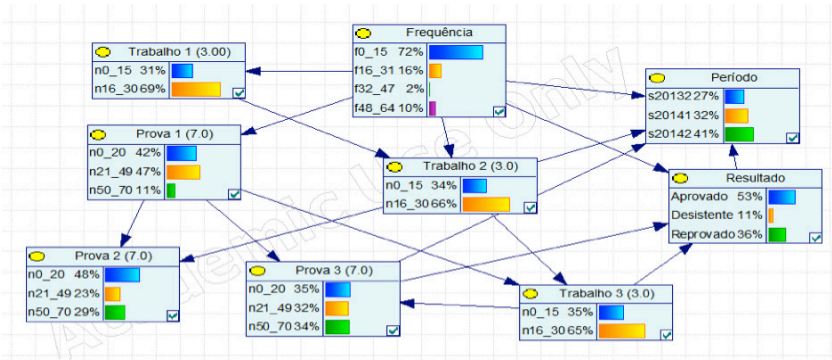

Figura 1. Rede Bayesiana gerada através do algoritmo *Greedy Thick Thinning*.

Realizando inferências nessa rede, foi selecionado 100% dos casos (barra azul) em que alunos não realizaram a entrega dos trabalhos 1, 2 e 3 ou obtiveram nota destes trabalhos até 1,5 pontos e a quantidade de faltas entre 0 e 15. Assim, percebeu-se que essa configuração representa 76% (barra verde) dos alunos reprovados e 12% dos alunos desistentes (barra laranja), como mostra a Figura 2.

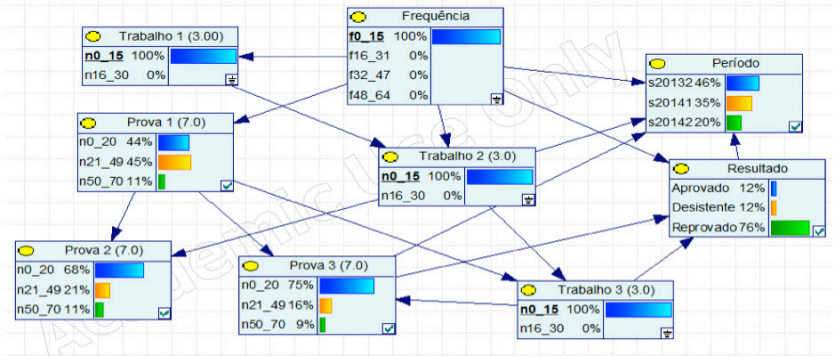

Figura 2. Inferência com nós Trabalho 1, Trabalho 2, Trabalho 3 e Frequência.

Analisando especificamente as notas da unidade 1, foi selecionado 100% dos alunos que tiraram nota no Trabalho 1 até 1,5 pontos e na Prova 1 até 2,0 pontos, e verificou-se que esta configuração resulta em uma chance de apenas 17% (barra azul do nó Resultado) de aprovação na disciplina, conforme mostra a Figura 3. A unidade 1 é a mais elementar e primordial para o entendimento dos conteúdos posteriores.

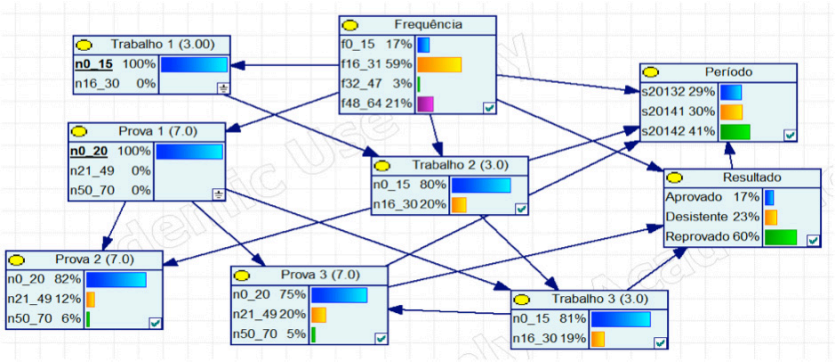

Figura 3. Inferência com nós Trabalho 1 e Prova 1.

4.4.2. Algoritmo *Bayesian Search*

A Figura 4 ilustra a rede gerada pelo algoritmo *Bayesian Search*. Esta se apresenta com 12 arcos e, em relação à rede gerada pelo *Greedy Thick Thinning*, possui menos conectividade.

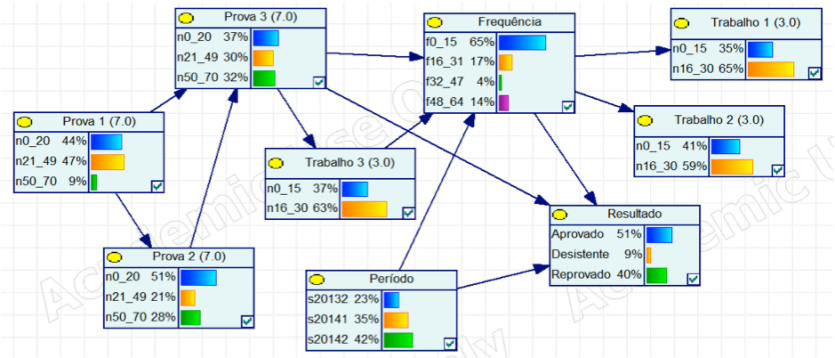

Figura 4. Rede Bayesiana gerada pelo algoritmo *Bayesian Search*.

Com o intuito de encontrar uma rede Bayesiana mais condizente com a realidade, foi utilizado um recurso da ferramenta GeNIe que permite ao usuário fornecer conhecimento prévio do domínio. Assim, as avaliações por unidade foram agrupadas

(trabalho + prova) e colocadas em na ordem temporal dos acontecimentos por unidade. Reforçou-se, também, por meio de uma ligação, o fato de que a frequência implica diretamente o resultado final de cada aluno na disciplina. A Figura 5 apresenta como essas informações foram fornecidas ao GeNIe 2.1.

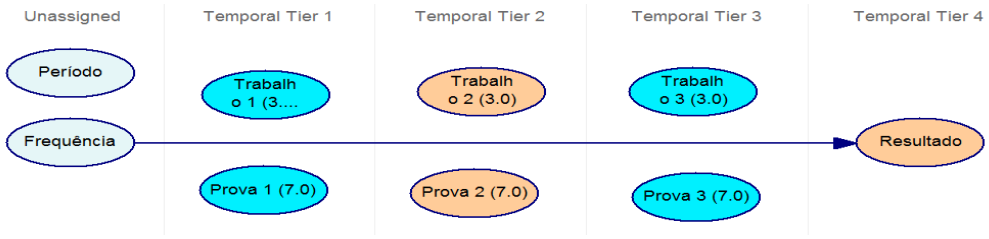

Figura 5. Fornecimento de Conhecimento do Domínio.

Assim, foi gerada uma nova rede pelo *Bayesian Search* com a inserção do conhecimento prévio do domínio, tendo somente 9 arcos. Realizando inferências (seleção de atributos) nesta rede, observou-se que os alunos que acompanharam toda a disciplina (frequência entre 0 e 15 faltas), no semestre letivo 2014.2 (barra verde do nó Período), obtendo notas boas em todos os trabalhos (100% dos alunos com notas entre 1,6 e 2,0 nos trabalhos), tiveram 90% de chance de serem aprovados (barra azul do nó Resultado), como mostra a Figura 6. Este semestre, sobretudo, foi extremamente difícil para os alunos porque houve uma longa greve dos professores, mas, ainda assim, o desempenho da disciplina foi considerado satisfatório.

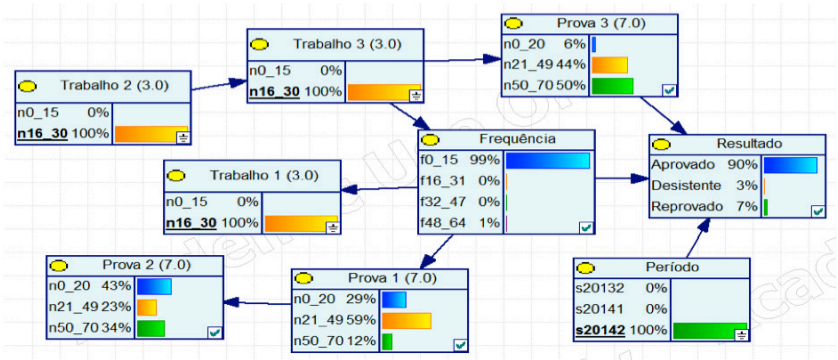

Figura 6. Inferência com nós Trabalho 1, Trabalho 2, Trabalho 3 e Período.

Analisando de maneira inversa, tentando compreender situações para a ocorrência de um aluno desistente na disciplina, pode-se observar a inferência realizada no nó Resultado com o valor desistente, conforme mostra a Figura 7.

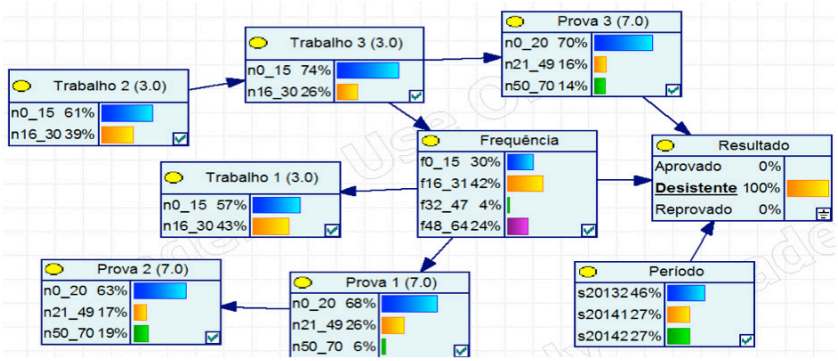

Figura 7. Inferência com o nó Resultado.

Neste caso, quando um aluno foi considerado desistente, observa-se dois aspectos: o primeiro é que somente 24% destes alunos nunca compareceram às aulas, tendo número alto de faltas (entre 48 e 64 horas), ou seja, o aluno desiste mesmo assistindo algumas aulas; e o segundo é que um aluno desistente tem 68% de chance de ter tirado nota entre 0 a 2,0 na prova 1, desmotivando-o desde o início da disciplina.

5. Considerações Finais

O presente trabalho teve o propósito de analisar o desempenho de alunos por meio de redes Bayesianas, além de reafirmar a necessidade e a importância da mineração dos dados educacionais. Diante dos resultados obtidos com a ferramenta GeNIe, conclui-se que a realização dos trabalhos (1, 2 e 3), mesmo que tenham pesos baixos (30%) em comparação com às provas (70%), em conjunto com a frequência influenciam fortemente a aprovação do aluno. Além disso, foi possível identificar, probabilisticamente, variáveis que influenciam e sinalizam o futuro do desempenho de alunos na disciplina de Lógica Aplicada à Computação no que diz respeito à aprovações, reprovações e desistências. Vale a pena repetir os experimentos realizados utilizando dados de mais outros semestres para consolidar os indícios preliminarmente identificados para que, em momentos futuros, os professores possam tomar decisões tentando minimizar o índice de reprovações e de desistências na disciplina.

Nesse contexto, para trabalhos futuros, pretende-se adicionar dados sociais e pessoais dos alunos à base de dados, para assim, melhor compreender as causas dos índices de reprovação e desistência. Além disso, aplicar outros algoritmos de aprendizado de estrutura, com diferentes parâmetros, para tentar encontrar a rede que seja mais próxima da realidade do domínio analisado.

Referências

Baker, R., Isotani, S. and Carvalho, Adriana. (2011) "Mineração de Dados Educacionais: Oportunidades para o Brasil", http://www.br-ie.org/pub/index.php/rbie/article/view/1301/1172, Fevereiro. In: Revista Brasileira de Informática na Educação. [S.l.], v. 19, n. 02, p. 03. ISSN 1414-5685.

BayesFusion, LCC. (2016) "GeNIe", http://www.bayesfusion.com/, Fevereiro.

Borges, J. P. V. and Padilha, T. P. P. (2005) "Modelagem do Processo de Aprendizado Colaborativo Através de Redes Bayesianas". In: VII Encontro De Estudantes De Informática Do Estado Do Tocantins, Palmas. Anais.

Costa, F. S. (2013) "Aprendizagem estrutural de redes Bayesianas pelo método de Monte Carlo e cadeias de Markov". 90 p. Monografia (Pós-Graduação em Ciência da Computação) - Universidade Federal de Santa Catarina, Florianópolis – SC.

Fayyad, U., Piatetsky-shapiro, G. and Smyth, P. (1996) "From data mining to knowledge discovery: An overview". In: Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, MIT, Cambridge, Massachusetts, and London, England, p.1-34.

Marques, Aramis Ferreira. (2014) "Aplicação de clusterização de dados na base de dados do zoneamento ecológico-econômico de Minas Gerais". 88 p. Monografia (Bacharelado em Sistemas da Informação) - Universidade Federal de Lavras, Lavras – MG.

RapdMiner. (2016) "RapdMiner", https://rapidminer.com/, Janeiro.