Modalidade do trabalho: Relatório técnico-científico
Evento: 2011 SIC - XIX Seminário de Iniciação Científica

DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS: APLICADO AO CASO DE ALUNOS COM SUBSÍDIO NO NÚCLEO DE ASSITÊNCIA ESTUDANTIL – NAE DA UNIVERSIDADE FEDERAL DO RIO GRANDE - FURG[1]

Wagner Gadêa Lorenz[2]; Leonardo Ramos Emmendorfer[3]; Adriano Velasque Werhli[4]

[1] Trabalho resultande de Monografia de Conclusão do Curso de Engenharia de Computação da Universidade Federal do Rio Grande - FURG

[2] Estudante do Curso de Engenharia de Computação do Centro de Ciências Computacionais – C3 da Universidade Federal do Rio Grande – FURG. E-mail: wagnerglorenz@gmail.com.

[3] Professor Doutor do Centro de Ciências Computacionais – C3 da Universidade Federal do Rio Grande – FURG, Orientador. E-mail: leonardo.emmendorfer@gmail.com.

[4] Professor Doutor do Centro de Ciências Computacionais – C3 da Universidade Federal do Rio Grande – FURG, Co-orientador. E-mail: werhli@gmail.com.

Resumo

O presente artigo tem como objetivo propor o uso de ferramentas e tecnologias para análise de bancos de dados, com o intuito de encontrar padrões e informações relevantes a novos conhecimentos. Utilizado conceitos e aplicações do processo *KDD* (*Knowledge Discovery in Databases),* também conhecido como descoberta de conhecimento em bases de dados. Dados da Universidade Federal do Rio Grande - FURG foram escolhidos intencionalmente de maneira a fornecer informações para a universidade sobre um modelo de classificação dos alunos com subsídio no Núcleo de Assistência Estudantil - NAE. Após a realização as principais etapas do processo de descoberta de conhecimento em bases de dados: Pré-processamento, Mineração de Dados e Pós-processamento. Baseado em dados de concessões anteriores obteve-se como saída um modelo do comportamento do aluno com bom desempenho versus aluno com baixo desempenho.

Palavras-chave: Mineração de Dados; Pré-processamento; Pós-processamento.

Introdução

O constante avanço na área da Tecnologia da Informação viabilizou o armazenamento de grandes e múltiplas bases de dados. A partir de vários tipos de tecnologias como internet, sistemas de gerenciadores de bancos de dados, sistemas de informação em geral, têm viabilizado uma grande proliferação de inúmeras bases de dados de natureza comercial, científica, governamental e administrativa. Com isso surge a necessidade de analisar e utilizar do melhor modo possível todo esse volume de dados disponível obtendo informações até o momento implícitas, segundo Goldschmidt e Passos (2005).

De acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996): KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados.

Segundo Goldschmidt e Passos (2005), a descoberta de conhecimento em bases de dados é multidisciplinar e, historicamente, origina-se de diversas áreas, dentre as quais podem

Modalidade do trabalho: Relatório técnico-científico
Evento: 2011 SIC - XIX Seminário de Iniciação Científica

ser destacadas: Estatística, Inteligência Computacional, Aprendizado de Máquina, Reconhecimento de Padrões e Banco de Dados.

Dados da Universidade Federal do Rio Grande foram escolhidos para que, ao realizar o projeto de pesquisa e aplicação, seja possível fornecer informações para a universidade sobre classificação e um modelo de escores dos alunos com subsídio em programas de apoio do Núcleo de Assistência Estudantil, tendo como variável resposta o comportamento do aluno com boa performance versus baixa performance, que hoje estão implícitos, para posterior análise e, assim, contribuir de alguma forma com a instituição.

Metodologia

A Figura 1, adaptada de Fayyad, Piatetsky-Shapiro e Smyth (1996), traz uma visão das etapas do processo de Descoberta de Conhecimento em Bases de Dados.

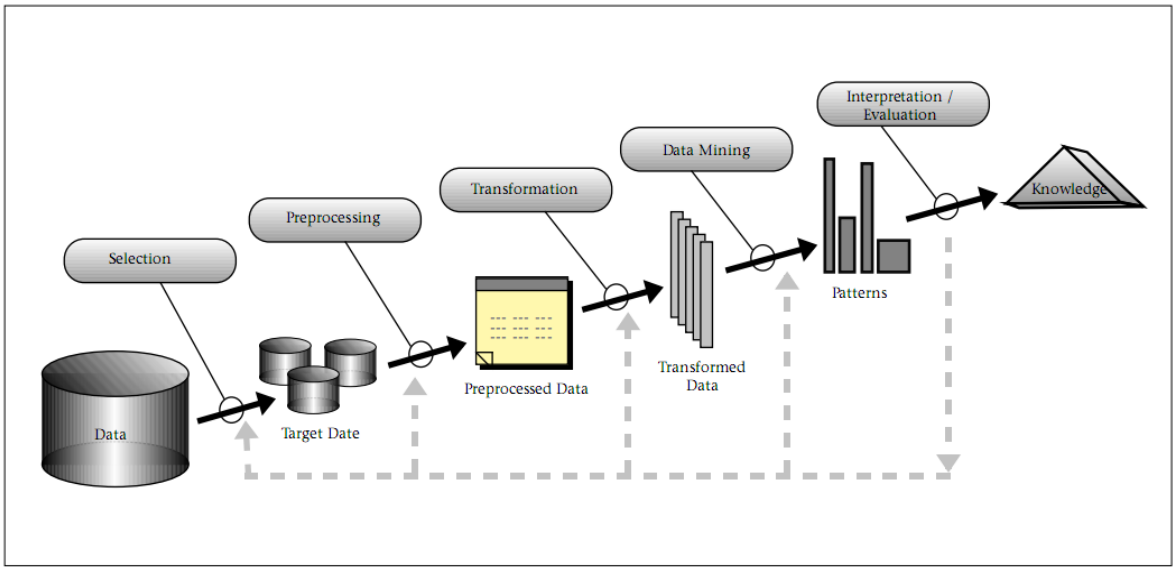

Figura 1 - Visão geral das etapas do processo de Descoberta de Conhecimento em Bases de Dados – KDD.

De acordo com o agrupamento proposto por Goldschmidt e Passos (2005), a etapa de pré-processamento tem as funções de captação, organização e tratamento dos dados, com o objetivo de preparar os dados para os algoritmos da etapa de Mineração de Dados.

Algumas das mais importantes atividades da etapa de pré-processamento de dados, de acordo com Goldschmidt e Passos (2005), Zhang, Zhang e Yang (2003), Tan, Steinbach e Kumar (2009) estão expostas a seguir:

- Seleções de Dados: foi utilizada a base de dados do NAE. De acordo com Goldschmidt e Passos (2005), quando os dados estão reunidos em uma mesma estrutura, a função de seleção de dados pode ter dois enfoques distintos: redução de dados horizontal e redução de dados vertical. No trabalho em questão foi utilizado a redução de dados horizontal, onde foi efetuada a eliminação direta de casos utilizando intruções da linguagem SQL, escolhendo casos para fazer parte ou não do conjunto de dados.

- Limpeza: o objetivo desta etapa é a eliminação de valores ausentes em conjuntos de dados segundo Goldschmidt e Passos (2005). Foi efetuada uma avaliação sobre a consistência das informações, correção de possíveis erros, eliminação de valores não pertencentes ao domínio e tratamento de valores ausentes ou redundantes.
- Enriquecimento: esta etapa serve para aumentar a base de dados utilizando outras fontes de informação de acordo com Elmasri e Navathe (2003). Foi utilizada a base de dados do Sistema de Informações Acadêmicas – SIA, para agregar informações à base já existente.
- Codificação: codificar significa transformar a natureza dos valores de um atributo de acordo com Goldschmidt e Passos (2005). Os dados foram codificados para atender às necessidades específicas dos algoritmos de mineração de dados.
- Construção de Atributos: onde são criadas novas colunas na tabela, de acordo com Soares (2007). Foi utilizado para criação de alguns tipos de atributos como: idade, área do curso.

A segunda etapa, Mineração de Dados temos à busca efetiva por conhecimentos úteis para o contexto da aplicação de KDD. Foi utilizado o *software* WEKA (*Waikato Enviroment for Knowledge Analysis*) para classificação, aplicando árvore de decisão para hierarquização dos dados. De acordo com Souza, Mattoso e Ebecken (1998) e Bogorny (2003) , árvore de decisão é a maneira mais simples de classificar exemplos em um número finito de classes.

A terceira etapa, Pós-processamento visa abranger o tratamento do conhecimento obtido através da Mineração de Dados.

Resultados e Discussão

Com a utilização do *software* WEKA, de acordo com Waikato (2008), WEKA é uma coleção de algoritmos de aprendizagem de máquina e ferramentas de pré-processamento de dados desenvolvida pela *University of Waikato*, Nova Zelândia. Com a utilização do algoritmo J4.8, de acordo com Witten e Frank (2005), o J4.8 é uma versão melhorada do popular C4.5. Sendo utilizados nos testes os parâmetros padrões do algoritmo J4.8. Também foi utilizada a validação cruzada 10 *folds*.

Com o atributo *nota* escolhido como classe para o algoritmo de classificação, apresentando os seguintes valores possíveis: '*BOA*' (para nota boa) e '*RUIM*' (para nota ruim). Em média, o índice de instâncias classificadas corretamente pelo algoritmo J4.8 da ferramenta WEKA, para os anos de 2009, 2010 e 2011, utilizando 6427 números de instâncias, é 70,3127% representam um bom nível de acerto segundo a literatura consultada. Consequentemente, a porcentagem de instâncias classificadas incorretamente pelo mesmo algoritmo foi de 29,6873%.

Na Tabela 1, encontram-se os atributos utilizados na ferramenta WEKA.

Tabela 1 - Atributos selecionados na ferramenta WEKA.

Atributo	**Referência**
area_curso	A área do Curso foi categorizada em EXATAS, SOC/HUMANAS e SAUDE.
Idade	A idade do aluno foi categorizada em menor de 20 anos, entre 21 e 25 anos e

	maior de 26 anos.
Sexo	O sexo do aluno foi categorizado em M ou F
Nota	A média das notas do aluno foi categorizada em acima ou igual a 7 como '*BOA'* e abaixo de 7 como '*RUIM'*.
Município	O município de origem do aluno foi categorizado em RIO GRANDE e OUTRAS.
ano_inscricao	O ano em que o aluno fez a inscrição em algum subsídio do NAE.
Subprograma	O subprograma do NAE que o aluno está inscrito, sendo eles: Moradia, Transporte, Alimentação, Bolsa Permanência e Pré-escola. Sendo categorizados em: '*M*', '*T*', '*A*', '*BP*' e '*PE*' respectivamente.

Após a seleção dos atributos, foi executada a etapa de D*ata Mining* na ferramenta WEKA somente com os atributos anteriores selecionados. A Figura 2 representa a árvore de decisão gerada pela ferramenta WEKA.

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

area_curso = EXATAS
|   idade = menor_20
|   |   sexo = M
|   |   |   municipio = RIO_GRANDE
|   |   |   |   subprograma = M: RUIM (9.0/3.0)
|   |   |   |   subprograma = T: BOA (90.0/42.0)
|   |   |   |   subprograma = A: BOA (84.0/39.0)
|   |   |   |   subprograma = BP: RUIM (34.0/12.0)
|   |   |   |   subprograma = PE: RUIM (0.0)
|   |   |   municipio = OUTRAS: BOA (67.0/24.0)
|   |   sexo = F
|   |   |   municipio = RIO_GRANDE: RUIM (164.0/65.0)
|   |   |   municipio = OUTRAS
|   |   |   |   subprograma = M: RUIM (18.0/6.0)
|   |   |   |   subprograma = T: BOA (30.0/13.0)
|   |   |   |   subprograma = A: BOA (43.0/18.0)
|   |   |   |   subprograma = BP: RUIM (35.0/14.0)
|   |   |   |   subprograma = PE: RUIM (0.0)
|   idade = entre_21_25: RUIM (1371.0/470.0)
|   idade = maior_26: RUIM (599.0/106.0)
area_curso = SOC/HUMANAS: BOA (3031.0/859.0)
area_curso = SAUDE
|   sexo = M
|   |   idade = menor_20: RUIM (23.0)
|   |   idade = entre_21_25: BOA (93.0/32.0)
|   |   idade = maior_26: RUIM (48.0/19.0)
|   sexo = F: BOA (688.0/165.0)

Number of Leaves  :     19

Size of the tree :  28
```

Figura 2 - Árvore de decisão gerada pela ferramenta WEKA.

Baseado em dados de concessões anteriores foi obtido como saída um modelo do comportamento do aluno com bom desempenho versus aluno com baixo desempenho.

Conclusões

O presente trabalho está em andamento, os resultados preliminares apresentados nesse resumo cumprem os primeiros objetivos, a descrição e apresentação da proposta.

Os dados obtidos serão apresentados para o Núcleo de Assistência Estudantil – NAE da Universidade Federal do Rio Grande - FURG, onde dependendo da avaliação e intenção de uso, serão criadas regras de classificação. Com a árvore de decisão gerada é possível extrair regras de classificação para o conjunto de entrada, a partir dos nós da árvore e do índice de acerto de classificação de cada nó folha.

Agradecimentos

A Pró-reitora de Assuntos Estudantis, Darlene Pereira por disponibilizar e permitir a utilização da base de dados do Núcleo de Assistência Estudantil – NAE.

A Pró-reitora de Graduação, Leila Costa Valle, por disponibilizar e permitir a utilização da base de dados do Sistema de Informações Acadêmicas – SIA.

Ao Diretor do Núcleo de Tecnologia da Informação - NTI, Marco Antônio Carou Leandro por disponibilizar os meios de obtenção dos dados.

Referências

BOGORNY, Vania. **Algoritmos e Ferramentas de Descoberta de Conhecimento em Bancos de Dados Geográficos**. Porto Alegre: PPGC da UFRGS. 2003.

ELMASRI, Ramez.; NAVATHE, Shamkant. B. ***Fundamentals of Database Systems*** **-** 4 ed. Pearson – Addison Wesley. 2003.

FAYYAD, Usama.; PIATETSKY-SHAPIRO, Gregory.; SMYTH, Padhraic. ***From Data Mining to Knowledge Discovery in Databases.*** *Ai Magazine*, 37-54. 1996.

GOLDSCHMIDT, Ronaldo.; PASSOS, Emmanuel. ***Data Mining*****: Um guia prático** - 2 ed. Rio de Janeiro: Campus. 2005.

SOARES, Jorge de. Abreu. **Pré-processamento em mineração de dados: um estudo comparativo em complementação.** PhD *thesis*, Universidade Federal do Rio de Janeiro, COPPE. 2007.

SOUZA, Mauro. S.; MATTOSO, Marta. L.; EBECKEN, Nelson. ***Data mining: a database perspective***. Rio de Janeiro: International Conference on Data Mining. 1998.

TAN, Pang-Ning.; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Data Mining: Mineração de Dados**. Rio de Janeiro: Ciência Moderna. 2009.

WAIKATO, University. ***WEKA 3: Data Mining Software in Java***. Nova Zelândia. 2008.

WITTEN, Ian. H.; FRANK, Eibe. ***Data Mining - Practical Machine Learning Tools and Techniques with Java Implementations.*** - 2 ed. Elsevier: Morgan Kaufmann Series. 2005

ZHANG, Shichao ; ZHANG, Chengqi.; YANG, Qiang. ***Data Preparation for Data Mining***. *Applied Artificial Intelligence*, 375-381. 2003.

Modalidade do trabalho: Relatório técnico-científico
Evento: 2011 SIC - XIX Seminário de Iniciação Científica