Técnicas de Mineração de Dados aplicado na Universidade Federal Rural do Semi-Árido (UFERSA) Campus Angicos[1*]

Julio Cartier Maia Gomes[1], Paulo Henrique de Morais[1], Cynthia Moreira Maia[1], Walter Martins Rodrigues[1]

[1]Universidade Federal Rural do Semi-Árido (UFERSA)
CEP 59515-000 – Angicos – RN – Brazil

Departamento de Ciências Exatas, Tecnológicas e Humanas – DCETH
Universidade Federal Rural do Semi-Árido (UFERSA) – Angicos, RN – Brazil

juliocartier@gmail.com, {paulomoraiis,cynthia-norte}@hotmail.com,walterm@ufersa.edu.br

Abstract. *The Information Technology improvements made possible the creation of several computational systems, for that one of the main objectives of the organization has been to store data. Data Mining emerged as a set of automation techniques for the exploration of large bulks of data, with the aim of discovering new patterns and relations that due to the large volume of data were difficult to be discovered before. Companies are becoming more and more accustomed to using this technology. The work analyzes a database equivalent to all the students of the Federal University Rural do Semi-Arido (UFERSA) Center of Angicos-RN, from the academic periods of 2012.1 to 2015.2, through these data a study is made.*

Resumo. *O avanço da Tecnologia da Informação possibilitou a criação de vários sistemas computacionais, com isso, um dos principais objetivos das organizações tem sido o de armazenar dados. A mineração de dados (*do inglês*, Data Mining) surgiu como um conjunto de técnicas automáticas para a exploração de grandes massas de dados, de forma a descobrir novos padrões e relações que, devido ao grande volume de dados antes eram dificilmente descobertos. As organizações estão cada vez mais se habituando ao uso dessa tecnologia. O trabalho analisa uma base de dados equivalente a todos os alunos da Universidade Federal Rural do Semi-Árido (UFERSA) Campus Angicos-RN, dos períodos letivos de 2012.1 ao 2015.2, através desses dados pôde-se realizar um estudo.*

1. Introdução

Desde o surgimento dos sistemas computacionais, um dos principais objetivos das organizações tem sido o de armazenar dados. Nas últimas décadas essa tendência ficou ainda mais evidente com a queda nos custos para a aquisição de hardware, tornando possível armazenar quantidades cada vez maiores de dados. Nesse contexto, foram desenvolvidas novas e mais complexas estruturas de armazenamento de dados, tais como: banco de dados, Data Warehouses, Bibliotecas Virtuais, Web e outras [Han & Kamber 2006].

As técnicas tradicionais de exploração de dados não são mais adequadas para tratar a grande maioria dos repositórios. Com a finalidade de solucionar este problema, foi proposta, no final da década de 80, a Mineração de dados (do inglês, *Data Mining*). A mineração de dados é uma das tecnologias mais promissoras da atualidade, podendo ser definida como um conjunto de técnicas automáticas de exploração de grandes massas de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertas [Amorim 2006].

Os analistas necessitavam de ferramentas capazes de auxiliá-los na análise de informações e consequentemente na mineração de dados. A partir dessa necessidade, um grupo de pesquisadores da Universidade de Waikato, na Nova Zelândia, desenvolveu (em java) um *software* livre para auxiliar na mineração de dados. O Weka (do inglês, *Waikato Enviroment for Knowledge Analysis*) [Pimenta et at. 2009], ao decorrer dos anos se consolidou como a ferramenta de *Data Mining* mais utilizada em ambiente acadêmico, apresentando como ponto forte a tarefa de classificação, sendo capaz também, de minerar regras de associação e regressão e agrupamento de dados.

Neste trabalho serão abordadas duas técnicas de mineração de dados com o auxílio do software Weka. Será realizada uma análise de uma base de dados com os alunos da Universidade Federal Rural do Semi-Árido Campus Angicos e seus principais atributos, relacionando e transformando os dados em informações e conhecimento.

2. Waikato Enviroment for Knowledge Analysis

O Weka é uma ferramenta de código aberto para mineração de dados, de interface amigável, que agrega um conjunto de algoritmos de classificação, regras de associação, regressão, pré-processamento e clustering, todos implementados em JAVA [Witten & Frank 2005]. Além disso, a aplicação Weka pode acessar dados oriundos de bancos de dados (via JDBC) ou através da chamada de arquivos de dados próprios [Pimenta et. at. 2009].

A ferramenta foi desenvolvida na Universidade de Waikato na Nova Zelândia, podendo ser definida, como uma coleção de algoritmos do tipo *machine learning* para tarefas de *data mining*. O Weka foi desenvolvido a partir de uma iniciativa da Fundação de Pesquisa, Ciência e Tecnologia do governo da Nova Zelândia e, vem sendo cada vez mais utilizado por possui características como, algoritmos para *Data Mining*, relativamente fáceis de usar, por proporcionar recursos flexíveis para experimentos e oferecer constantes atualizações, vale frisar, que muitos desses algoritmos incorporam conceitos de inteligência artificial [Markov & Russell 2006].

O Weka dispõe de diversas técnicas de mineração de dados e pode ser instalado em diferentes sistemas operacionais tais como: Windows Mac OS e Linux, sua distribuição se dá de forma livre por meio da Internet, podendo ser modificado e redistribuído de acordo com os termos da GNU (do inglês, *General Public License*) [Weka 2017].

3. Algoritmos Bagging e Boosting

Um dos métodos mais populares de aprendizagem por conjunto de classificação é conhecido como *bagging*, um acrônimo para *bootstrap e o boosting*. O *bagging* envolve a manipulação dos dados em conjuntos de treinamento, enquanto que outros métodos

podem ser omitidas desse conjunto. Para a geração de diferentes classificadores, os dados devem ser combinados em classificadores compostos, cuja classificação será obtida através da votação dos classificadores. Diversas pesquisas mostram que o método baseado em conjunto de classificador geralmente leva a ter melhorias significativas em uma série de problemas de aprendizagem: [Galar & Fernández 2012]. Devido ao fato da amostragem ser feita com substituição, algumas instâncias podem aparecer diversas vezes no mesmo conjunto de treinamento.

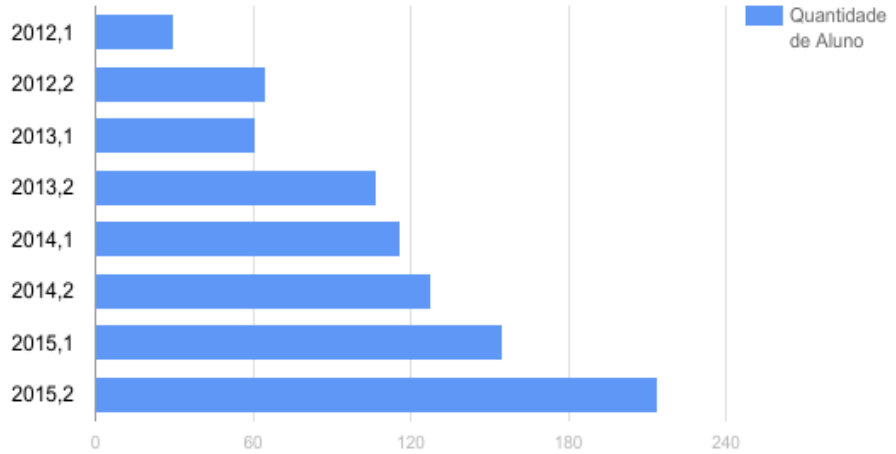
O *bagging* melhora o erro de generalização reduzindo a variância dos classificadores de base. O desempenho do algoritmo depende da estabilidade do classificador de base e se o classificador base for instável, o *bagging* auxilia a reduzir os erros associados às flutuações aleatórios nos dados de treinamento [Tan & Steinbach & Kumar 2009].

O método *boosting* é um procedimento iterativo usado para alterar adaptativamente a distribuição de exemplos de treinamento, de modo, que os classificadores de base enfoquem exemplos que sejam difíceis de classificar. Diferente do *bagging*, o *boosting* também atribui um peso a cada exemplo de treinamento podendo ser usado como uma distribuição de amostra para desenhar um conjunto de amostras de *bootstrap*. O *bootstrap* aborda os registros de treinamento que são amostrados como substituição a partir dos dados originais. O método *boosting* também pode ser usado pelo classificador de base para descobrir um modelo que tenha tendência na direção de exemplos de peso mais altos [Tan & Steinbach & Kumar, 2009].

4. Base de Dados

A Universidade Federal Rural do Semi-Árido (UFERSA) é uma das universidades que aderiu ao Programa de Reestruturação e Expansão das instituições Federais de Ensino. Em função disto, foi implantado o Campus na cidade de Angicos, localizada na Região Central do Estado do Rio Grande do Norte em 2009, até junho de 2016, contando com 876 alunos ativos na instituição como podemos ver o Gráfico 1 no período de 2012.1 à 2015.2.

Tabela 1. Alunos Ativos no período de 2012.1 até 2015.2



Para realização da pesquisa foram extraídas inicialmente algumas variáveis de conteúdos relevantes que podem ser úteis para descrever as características. Para tanto, foram selecionados os seguintes campos da tabela dos alunos ativos do Campus Angicos: Curso, Quantidade de Períodos Cursados, Status do Discente, Semestre de Entrada, Semestre de Saída, Tipo de Saída, Aprovações, Reprovações, IRA (Índice de Rendimento Acadêmico) e IEA (Índice de Eficiência Acadêmica).

5. Análise do resultado

Foram aplicados os algoritmos *bagging* e *boosting*, levando em conta todos os alunos com matrículas ativas, trancadas ou canceladas na Universidade Federal Rural do Semi-Árido (UFERSA) Campus Angicos com o grau de confiabilidade no total de instâncias que conseguiu se classificar ou não classificar corretamente como segue na Tabela 1.

Tabela 1: Instâncias Classificadas Corretamente ou Não Corretamente

Algoritmos	Instâncias Classificadas Corretamente (%)	Instâncias Não Classificadas Corretamente(%)
Bagging	68,75%	31,24%
Boosting	88,29%	11,70%

Tabela 2: Alunos divididos pelo o seu status de matrícula

Status dos Alunos	Alunos %	Quant. de Alunos
Ativos	59,5%	876
Ativos-Graduando	3,2%	48
Cancelados	28,9%	424
Concluídos	1,9%	93
Trancados	6,3%	28

Com a execução dos dois algoritmos no Software Weka aplicando à base de dados, foi gerada a matriz de confusão, à qual apresentou a classe de status dos alunos, como podemos observar na Tabela 2, em que 424 dos alunos do Campus Angicos têm a matrícula cancelada. Os alunos ativos equivalem a 876 com a maior proporção nos períodos de 2015.1 e 2015.2, assim que ingressaram no Campus. Em relação aos alunos que efetuaram trancamento do curso temos um numérico de 28 trancamentos. Dentre os 1479 alunos do período em discussão, só concluíram a graduação 93 alunos e, 48 alunos estão ativos cursando o segundo ciclo, que é o caso dos que concluem o curso de Ciência e Tecnologia e iniciam as engenharias.

6. Conclusão

Este trabalho visou explorar os algoritmos de mineração de dados bagging e boosting na Universidade Federal Rural do Semi-Árido (UFERSA) Campus Angicos, ao analisar os dados foi possível ter convicção de que os valores encontrados são de extrema

importância aos alunos do Campus. Uma vez, que os resultados obtidos nos apresentam a porcentagem dos alunos que cancelam as matrículas, os ativos, os que efetuam trancamentos, os que concluíram e os que vão para o segundo ciclo (engenharia).

Após a análise dos dados, constata-se que nosso trabalho é relevante, pois o mesmo proporcionou uma contribuição significativa a toda comunidade acadêmica, apresentando informações que levariam bastante tempo para serem processadas se não usássemos a ferramenta WEKA, logo, estes algoritmos utilizados no processo de mineração de dados permitem que os usuários transformem dados em informação de forma rápida e eficaz.

Referências

Amorim, T. (2006) Conceitos, técnicas, ferramentas e aplicações de Mineração de dados para gerar conhecimento a partir de bases de dados. Universidade Federal de Pernambuco.

Galar, M. Fernández, A. (2012) Barrencecha, E. Bustince, H, Herrera, F.: A Review on Ensembles for the Class Imbalance Problem: Bagging -, Boosting-, and, Hybrid-Based Approaches. Vol 42, No. 4, July.

Han, J Kamber, M. (2006) Data Mining: Concepts and Techniques. Elsevier.

Markov, Z. Russell, I. (2006) An Introduction to the WEKA Data Mining System. Proceedings of the 11th annual SIGCSE conference on Innovation and technology in computer science education. P. 367 - 368. Bologna, Italy.

Pimenta, A. Valentim, P. Santos, D. Neto M. (2009) Weka-G: Mineração de Dados Paralela em Grandes Computacionais: Revista Sistemas de Informação da FSMA, nº4. pp 2-9.

Tan, P. Steinbach, M. Kumar, V. (2009) Introdução ao Data Mining: Mineração de Dados. Rio de Janeiro: Editora Ciência Moderna Ltda.

Witten, I. H. Frank, E. (2005) “Data Mining: Practical machine learning tools and techniques”. 2a edição. Morgan Kaufmann, São Francisco.

Weka (2017) Data Mning Software in Java The University of Waikato,disponível em: <http://www.cs.waikato.ac.nz/ml/weka/> Acesso: 10 de fevereiro