Uma análise comparativa sobre ferramentas de mineração de dados adotadas na academia e na indústria

Comparative analysis of data mining tools used in academy and in industry

Thiago Affonso de M. N. Viana
tamnv@recife.ifpe.edu.br
Instituto Federal de Pernambuco

Resumo

Hoje em dia, existe uma grande quantidade de dados sendo produzidos através da utilização de sistemas de informação. Todos os dias, muitas informações são armazenadas em diversos bancos de dados, principalmente corporativos. Assim sendo, o processo de mineração de dados vem sendo amplamente difundido e utilizado, seja na academia ou na indústria, como um método de auxílio à tomada de decisão em áreas como medicina, meteorologia, áreas financeiras, entre outras. O uso de mineração de dados mostra-se eficiente quando se deseja obter informações que, a princípio, não são perceptíveis, no entanto, através de combinações promovidas nos dados, podem ser descobertas e usadas como fontes estratégicas para o apoio na tomada de decisões. Todavia, esse processo não é viável se não for realizado de maneira automatizada, e para isso, existem ferramentas que utilizam a mineração de dados para gerar resultados, tornando o processo de descoberta de conhecimento e mineração de dados mais rápido e eficiente. O trabalho aqui proposto descreve um estudo comparativo entre três ferramentas de mineração de dados, as quais foram escolhidas após uma pesquisa de uso na academia e na indústria da cidade do Recife-PE. Essas ferramentas foram analisadas sob duas perspectivas: desempenho, medido a partir do tempo necessário para o processamento dos dados de cada uma das ferramentas; e precisão, estudada a partir do percentual de acertos e erros gerados nos dados de saída das ferramentas.

Palavras-chave: Mineração de dados. Inteligência artificial. Análise comparativa.

Abstract

Today there is a large amount of data being produced through the use of information systems. Every day a lot of information is stored in various databases, especially corporate ones. Thus, the process of data mining, has been widely used, either in academia or industry, as a method to help in decision support in areas such as medicine, meteorology, financial, among others. The use of data mining has proved his effectiveness when wanting to obtain informations that, in principle, are not noticeable, however, promoted by combining the data can be discovered and used as sources to support strategic decision making. However, this process is not viable if not performed in an automated manner, and for that there are tools that use data mining to generate results, making the process of knowledge discovery and data mining faster and more efficient. The present work describes a comparative study of three data mining tools, which were chosen after a research of use in the academia and industry in the city of Recife / PE. These tools were analyzed on two perspectives: performance, measured from the time required for processing the data from each one of the tools and precision, studied from the percentage of right and wrong answers generated in the output of the tools.

Keywords: Data mining. Artificial intelligence. Comparative analysis.

1 Introdução

Com o crescimento da automatização de manipulação de informações através dos sistemas computacionais, os dados transitados pelo mundo também cresceram rapidamente. Com o custo de armazenamento cada vez mais barato e seguro, existem muitos dados guardados que estão disponíveis e podem ser mais bem aproveitados através de técnicas, ferramentas de análise e busca de informações (LIMA, 2009).

Os benefícios da utilização de uma busca de informações através da mineração de dados, processo que analisa os dados visando à descoberta de padrões novos e potencialmente úteis (FAYYAD *et al*., 1996), é algo que podemos considerar de grande valia. Existe uma quantidade alta de informações nos vários bancos de dados existentes que podem ser importantes e, por muitas vezes, acabam passando despercebidas, servindo apenas a um fim específico em nível operacional de uma empresa, não sendo aproveitadas em níveis estratégicos (BENICASA & PAIXAO, 2006).

Os benefícios advindos das técnicas de mineração de dados se tornam ainda mais vantajosos ao se utilizar ferramentas que automatizem a busca, a análise e o processamento dos dados. Assim sendo, a análise das características existentes em várias ferramentas pode fornecer uma resposta mais precisa e confiável, apoiando a escolha adequada da ferramenta a ser utilizada.

Este artigo se propõe a realizar um estudo comparativo sistemático entre as ferramentas de mineração de dados usadas na academia e na indústria. Elas foram analisadas a partir de dois importantes itens: desempenho, estudado a partir do tempo, em segundos, necessário à geração da resposta esperada; e precisão, estudada a partir do percentual de acertos e erros gerados nos dados de saída das ferramentas.

O artigo está estruturado da seguinte maneira: a Seção 2 discorre sobre os assuntos de mineração de dados, os quais são necessários para contextualizar os temas abordados neste trabalho. Em seguida, a Seção 3 demonstra a metodologia seguida para o estudo. A Seção 4 apresenta os resultados obtidos, bem como a sua discussão. Por fim, a Seção 5 apresenta as considerações finais, bem como as conclusões advindas dos resultados deste trabalho.

2 Mineração de dados

A técnica de mineração de dados (GOLDSHMIDT & PASSOS, 2005) surgiu com o processo de descoberta de conhecimento (BENICASA & PAIXAO, 2006), que busca, através de métodos automáticos baseados em estatística, a extração de conhecimento de alto nível, partindo de bases de dados reais. O processo de mineração de dados visa analisar os dados no intuito da descoberta de padrões novos, válidos e potencialmente úteis. Nesse contexto, são definidas quatro técnicas de mineração de dados: classificação, agrupamento, regressão e associação (HAN & KAMBER, 2001).

Através das técnicas de classificação (HAN & KAMBER, 2001) e de agrupamento (CÔRTES *et al*., 2002), é possível definir a qual grupo pertence um elemento qualquer. Nesse contexto, diversas áreas de conhecimento podem ser beneficiadas, como a área da medicina, classificando, por exemplo, que tipo de doença é mais característica em pessoas com sintomas X, Y e Z, baseado em casos anteriores; ou o campo da meteorologia, obtendo resposta sobre a possibilidade de um dia com características A, B e C ser chuvoso ou ensolarado.

Por outro lado, as técnicas de regressão (LENZ, 2009) são capazes de realizar previsão sobre valores numéricos, permitindo, por exemplo, a previsão de cotações de ações na bolsa de valores (AMORIM *et al*., 2010) ou, ainda, a faixa de crédito a se disponibilizar para um determinado cliente (FERNANDES & JUNIOR, 2007).

Por fim, as técnicas de associação permitem a criação de regras com as conexões das informações de uma base de dados. Isso é possível a partir de formações lógicas, tais como "*se X ocorre em um dado contexto, então, Y ocorre*". Um exemplo de aplicação dessa técnica é a identificação dos padrões que aparecem em uma base de dados sobre clientes que são bons ou maus pagadores. Nesse caso, é possível a criação de regras como: "*se o cliente é casado, então, é bom pagador*", "*se o cliente possui renda maior que R$ 500, então, é bom pagador*", ou ainda "*se cliente possui três filhos, então, é mau pagador*".

3 Metodologia

O estudo comparativo se iniciou a partir da definição de quais ferramentas seriam consideradas na avaliação. Foi realizada uma pesquisa na academia e na indústria para definir quais são as ferramentas de mineração de dados mais utilizadas no contexto local da cidade de Recife-PE. Para a academia, foram selecionadas quatro instituições de ensino superior, sendo três públicas (UFPE, UPE e IFPE) e uma privada (AESO). A escolha dessas instituições se deveu ao melhor acesso da equipe envolvida neste trabalho a elas. Para a indústria, foram selecionadas quatro empresas inseridas no Porto Digital da cidade do Recife.

A escolha foi dada após o levantamento de quais atuam no desenvolvimento de sistemas e utilizam técnicas de mineração de dados. Por questões particulares, elas não permitiram a publicação de seus nomes no presente artigo,

passando a serem referenciadas apenas como Empresa 1, Empresa 2, Empresa 3 e Empresa 4. O resultado da consulta pode ser visualizado na Tabela 1.

A partir desse resultado, foram escolhidas três ferramentas como alvo de estudo: Tanagra (RAKOTOMALALA, 2005), Matlab (MATHWORKS, 2011) e Weka (HALL *et al*., 2009).

Tabela 1: Resultado da consulta sobre ferramentas de mineração de dados mais usadas na academia e na indústria no contexto de Recife-PE.

Instituição	Ferramentas Usadas
UFPE	MATLAB, Weka
IFPE	Weka
UPE	Weka, MATLAB, Tanagra
AESO	WEKA
Empresa 1	MATLAB, Weka
Empresa 2	MATLAB
Empresa 3	MATLAB, Weka
Empresa 4	MATLAB

3.1 Ferramentas envolvidas no estudo

Esta subseção provê uma breve explicação das ferramentas de mineração de dados que foram utilizadas no estudo definido por este artigo: Tanagra, Matlab e Weka.

3.1.1 Tanagra

Tanagra é um *software* gratuito para mineração de dados, usado normalmente para fins acadêmicos e de pesquisa na área de mineração de dados. A partir da análise exploratória destes, propõe vários métodos. O Tanagra possui alguns algoritmos de agrupamento, análise fatorial, regras de associação, entre outros. É um projeto de código aberto, acessível e pode ser baixado e alterado caso se concorde com os termos da licença de distribuição, mas muitos pesquisadores utilizam esse código já pronto para inserir modificações e, com isso, continuar os experimentos. O principal objetivo do Tanagra é fornecer aos pesquisadores uma ferramenta gratuita que possa auxiliar na análise de dados reais ou sintéticos.

3.1.2 MATLAB

O MATLAB é um *software* comercial, pago e robusto, voltado para os cálculos numéricos e oferece um alto desempenho na produção de seus resultados. Além de cálculos numéricos, essa ferramenta também serve para a visualização e análise de dados.

Esse *software* permite exportar, em forma de gráfico ou relatórios completos, os resultados apresentados, fazendo com que a importação para programas como Word ou PowerPoint seja viável. O código MATLAB também pode ser publicado automaticamente em HTML, Word, Latex e outros formatos. Algumas importantes características podem ser notadas no MATLAB, como a linguagem de alto nível para computação técnica, as ferramentas interativas para a resolução de problemas, o *design* e a exploração interativa, e as ferramentas para a construção de interfaces gráficas.

3.1.3 WEKA

O Weka é uma coleção de algoritmos de aprendizado de máquina para mineração de dados. É comumente usada na academia para pesquisas e aplicações de estudos. Hoje, é pertencente ao pacote Pentaho (BOUMAN & DONGEN, 2009). A ferramenta trabalha com dados no formato .csv, mas possui um formato próprio, em que é possível, através de metadados, delimitar informações e realizar um pré-processamento destes. A ferramenta possui uma interface fácil e conta com um conjunto de dados para testes, pesquisas e estudos. Além de ser uma ferramenta gratuita, pode ser usada para fins comerciais e, como pôde ser visto na Tabela 1, duas das quatro empresas a utilizam com esse fim.

3.2 Bases de dados utilizadas no estudo

Para os testes do estudo comparativo, foram escolhidas duas bases de dados consideradas clássicas na academia: Iris e Wine, que são disponibilizadas e podem ser acessadas a partir do *site* UCI Machine Learning Repository (FRANK & ASUNCION, 2010).

- **Iris:** É uma base de dados de espécies de flores da família das Iridáceas chamada Iris. Existem três classes nesse banco de dados: a Iris-setosa, a Iris-versicolour e a Iris-virginica, e possui para cada classe 50 instâncias, totalizando 150 instâncias. Cada classe possui 4 atributos: o comprimento da sépala, a largura da sépala, o comprimento da pétala e a largura da pétala.
- **Wine:** É uma base de dados utilizada para determinar a origem dos vinhos a partir de uma análise química. A análise foi realizada na Itália, em três cultivos diferentes. Existem 59 exemplos do primeiro cultivo, 71 do segundo e 48 do terceiro, totalizando 178 exemplos na base de dados. Cada um deles possui 13 atributos.

3.3 Execução dos testes de desempenho e precisão

Objetivando realizar uma análise comparativa entre as ferramentas levantadas neste estudo, foram escolhidos dois parâmetros de comparação: desempenho, baseado no tempo necessário para o processamento dos dados; e precisão, baseado no percentual de acertos e erros gerados nos dados de saída das ferramentas. Utilizando-se das bases de dados citadas na Seção 3.2, cada uma das ferramentas foi utilizada para a mineração de dados nas quatro técnicas existentes, descritas na Seção 2: classificação, agrupamento, regressão e associação.

Os testes foram executados em um mesmo computador, para não haver divergências de resultados por parte de configuração de *hardware*. Durante o processamento dos dados, foram coletados, através de um mesmo cronômetro, os tempos de execução, dando origem aos valores de desempenho relatados na Seção 4. Após o processamento dos dados, os resultados gerados por cada uma das ferramentas foi comparado ao resultado ideal esperado. Assim, foi gerado um percentual de erro e de acerto para cada ferramenta na aplicação de cada técnica. Os resultados e a discussão estão relatados a seguir.

4 Análise dos resultados

Após a aplicação de algoritmos de cada uma das técnicas de mineração de dados (classificação, agrupamento, regressão e associação) em cada uma das ferramentas, foram obtidos os resultados apresentados nas tabelas que se seguem.

Na Tabela 2 e na Tabela 3, é possível observar os resultados obtidos nas três ferramentas para os algoritmos de classificação na base de dados Iris (Tabela 1) e na base de dados Wine (Tabela 2). É possível observar que, em relação ao desempenho, as três ferramentas apresentaram um comportamento similar, não havendo diferença significativa entre elas, destacando-se a ferramenta Weka como a de melhor desempenho, apesar de ser uma diferença realmente muito pequena. Já no quesito precisão, é possível observar que o MATLAB se destacou dos demais, apresentando 98% de acerto para a base Iris e 92% para a base Wine. O Tanagra e o Weka obtiveram uma precisão muito similar entre si, sendo o Weka um pouco melhor.

Tabela 2: Resultado do processamento da base de dados Iris com algoritmos de classificação.

Ferramenta	Desempenho (segundos)	Dados de saída corretos	Dados de saída incorretos
Tanagra	0,19s	91%	9%
MATLAB	0,16s	98%	2%
Weka	0,15s	94%	6%

Tabela 3: Resultado do processamento da base de dados Wine com algoritmos de classificação.

Ferramenta	Desempenho (segundos)	Dados de saída corretos	Dados de saída incorretos
Tanagra	0,19s	80%	20%
MATLAB	0,16s	92%	8%
Weka	0,16s	82%	18%

Na Tabela 4 e na Tabela 5, é possível observar os resultados obtidos nas três ferramentas para os algoritmos de agrupamento na base de dados Iris (Tabela 4) e na base de dados Wine (Tabela 5). É possível observar que, em relação ao desempenho, o MATLAB se destacou bastante dos demais, executando os seus algoritmos em um tempo 39% melhor ao do Tanagra e 36% melhor ao do Weka. Nesse mesmo quesito, o Tanagra e o Weka apresentaram resultados similares, com um leve melhor desempenho por parte do Weka. Já no que diz respeito à precisão, é possível observar que o MATLAB e o Weka alcançaram um ótimo resultado na base de dados Iris, enquanto o Tanagra possuiu um resultado bem inferior. Já na base de dados Wine, o MATLAB obteve um ótimo resultado, com 100% de acerto, enquanto o Weka apresentou um bom resultado, com 80% de acerto, e o Tanagra ficou com 60%, um resultado bem inferior.

Tabela 4: Resultado do processamento da base de dados Iris com algoritmos de agrupamento.

Ferramenta	Desempenho (segundos)	Dados de saída corretos	Dados de saída incorretos
Tanagra	0,43s	75%	25%
MATLAB	0,26s	100%	0%
Weka	0,41s	100%	0%

Tabela 5: Resultado do processamento da base de dados Wine com algoritmos de agrupamento.

Ferramenta	Desempenho (segundos)	Dados de saída corretos	Dados de saída incorretos
Tanagra	0,43s	60%	40%
MATLAB	0,26s	100%	0%
Weka	0,41s	80%	20%

Na Tabela 6 e na Tabela 7, é possível observar os resultados obtidos nas três ferramentas para os algoritmos de regressão na base de dados Iris (Tabela 6) e na base de dados Wine (Tabela 7). É possível observar que, em relação ao desempenho, o MATLAB se destacou bastante dos demais, executando os seus algoritmos em um tempo 36% melhor ao do Tanagra e 12% melhor ao do Weka. Nesse mesmo quesito, o Weka apresentou um melhor desempenho, comparado ao Tanagra. Já no quesito precisão, é possível observar que o MATLAB se destacou dos demais, possuindo 90% de acerto para a base Iris e Wine. O Tanagra e o Weka obtiveram uma precisão muito similar entre si na base de dados Iris, sendo o

Weka um pouco melhor. Já na base de dados Wine, o desempenho do Weka foi significativamente superior ao do Tanagra.

Tabela 6: Resultado do processamento da base de dados Iris com algoritmos de regressão.

Ferramenta	Desempenho (segundos)	Dados de saída corretos	Dados de saída incorretos
Tanagra	0,55s	80%	20%
MATLAB	0,35s	90%	10%
Weka	0,40s	83%	17%

Tabela 7: Resultado do processamento da base de dados Wine com algoritmos de regressão.

Ferramenta	Desempenho (segundos)	Dados de saída corretos	Dados de saída incorretos
Tanagra	0,55s	70%	30%
MATLAB	0,35s	90%	10%
Weka	0,40s	79%	21%

Na Tabela 8 e na Tabela 9, é possível observar os resultados obtidos nas três ferramentas para os algoritmos de associação na base de dados Iris (Tabela 8) e na base de dados Wine (Tabela 9). É possível observar que, em relação ao desempenho, o Weka e o MATLAB se destacaram em relação ao Tanagra, sendo o Weka um pouco melhor em comparação ao MATLAB. Já no quesito precisão, é possível observar que o MATLAB se destacou dos demais, possuindo 90% de acerto para a base Iris e 85% de acerto para a base Wine. É possível, ainda, observar que o Weka obteve uma precisão bem maior que o Tanagra nas duas bases de dados.

Tabela 8: Resultado do processamento da base de dados Iris com algoritmos de associação.

Ferramenta	Desempenho (segundos)	Dados de saída corretos	Dados de saída incorretos
Tanagra	0,50s	75%	25%
MATLAB	0,35s	90%	10%
Weka	0,30s	83%	17%

Tabela 9: Resultado do processamento da base de dados Wine com algoritmos de associação.

Ferramenta	Desempenho (segundos)	Dados de saída corretos	Dados de saída incorretos
Tanagra	0,50s	65%	35%
MATLAB	0,35s	85%	15%
Weka	0,30s	75%	25%

Após a análise dos resultados, um valor médio foi obtido a partir dos dados gerados pela execução das ferramentas nas duas bases de dados usadas neste trabalho. A Tabela 10 demonstra o valor médio dos desempenhos de cada uma das ferramentas. Nela, estão destacados os melhores valores para cada técnica de mineração de dados. É possível observar que o MATLAB apresentou o melhor desempenho nas técnicas de agrupamento e regressão, enquanto o Weka apresentou melhor desempenho nas técnicas de classificação e associação, embora, nesses casos, o valor apresentado pelo Weka não tenha sido significativamente melhor se comparado ao MATLAB.

A Tabela 11 demonstra o valor médio da precisão, baseado nos percentuais dos dados de saída corretos de cada uma das ferramentas, sendo destacados os melhores valores para cada técnica. É possível observar que o MATLAB apresentou, em todas as técnicas, um resultado significativamente melhor. Contudo, os valores apresentados pelo Weka podem ser considerados bastante satisfatórios, possuindo uma média de 84% de acerto nas quatro técnicas de mineração de dados.

Tabela 10: Resultado médio de desempenho das ferramentas nas duas bases de dados.

Ferramenta	Classificação	Agrupamento	Regressão	Associação
Tanagra	0,19	0,43	0,55	0,50
MATLAB	0,16	***0,26***	***0,35***	0,35
Weka	***0,15***	0,41	0,40	***0,30***

Tabela 11: Resultado médio de precisão, baseado apenas nos percentuais dos dados de saída corretos das ferramentas nas duas bases de dados.

Ferramenta	Classificação	Agrupamento	Regressão	Associação
Tanagra	85,5%	67,5%	75%	70%
MATLAB	***95%***	***100%***	***90%***	***87,5%***
Weka	88%	90%	81%	79%

Podemos, com estes dados, concluir que, tratando-se de bases de dados como as apresentadas por este trabalho, o MATLAB é uma ferramenta que possui grandes qualidades de desempenho e precisão. O seu reconhecimento também pode ser observado, conforme descrito na Tabela 1, com o seu uso sendo constatado em duas das quatro instituições de ensino e em todas as quatro empresas utilizadas como fontes deste trabalho.

Dentre as várias qualidades apresentadas pelo MATLAB, pode-se salientar o alto número de algoritmos presentes na ferramenta, além do fato de permitir exportar os resultados apresentados em forma de gráfico ou relatórios completos, fazendo com que a importação para programas como Word ou PowerPoint seja viável. Algumas importantes características podem ser notadas ainda: a linguagem de alto nível para computação técnica, as ferramentas interativas para a resolução de problemas, o *design* e a exploração interativa, e as ferramentas para a construção de interfaces gráficas.

Contudo, para utilizar o MATLAB, é necessário realizar o pagamento da sua licença, o que, em muitos casos, pode ser inviável para algumas empresas e mais ainda para instituições de ensino e pesquisa. Sendo assim, os resultados apresentados demonstram que, para bases de dados simples e que representem comuns exemplos acadêmicos, tais como são as bases usadas neste estudo, o Weka apresenta resultados bastante proveitosos e viáveis para a aplicação das técnicas de mineração de dados. Além de ser gratuito e ter foco acadêmico, foi possível constatar neste estudo, conforme a Tabela 1, que todas as quatro instituições de ensino e duas das quatro empresas utilizadas como fonte de pesquisa neste trabalho usam o Weka.

5 Conclusão

Este trabalho procurou realizar um estudo comparativo entre ferramentas de mineração de dados utilizadas na academia e na indústria na cidade de Recife-PE. Para tal, foi realizada uma pesquisa de utilização em quatro instituições de ensino e quatro empresas da cidade. Após o levantamento das ferramentas, elas foram avaliadas sob duas perspectivas: desempenho, baseado no tempo necessário para a execução das técnicas de mineração de dados; e precisão, baseada no percentual de acerto e erro gerado pela comparação entre a saída gerada pelas ferramentas e a saída correta esperada. Todos os testes foram gerados usando duas bases de dados comumente empregadas como exemplos acadêmicos.

Após a coleta e análise dos resultados, foi observado que a ferramenta MATLAB apresentou o melhor desempenho e precisão na maioria das técnicas de mineração de dados. Corroborando com este resultado, pode-se observar que as quatro empresas e duas das quatro instituições de ensino usadas como fonte deste estudo fazem uso do MATLAB. Sendo assim, ele é uma ótima opção para aqueles que tenham interesse em adquirir a sua licença. Caso contrário, o estudo demonstrou que, para bases simples, especialmente as utilizadas em exemplos acadêmicos, o Weka apresenta um comportamento satisfatório, de modo que duas das quatro empresas e todas as instituições de ensino usadas como fonte deste estudo fazem uso dele.

Este estudo é um pequeno exemplo de comparação possível entre ferramentas de mineração de dados. Tendo em vista que é uma tecnologia em ascensão, trabalhos que venham somar são bem-vindos e suas conclusões servem de ajuda para o apoio a decisões e pesquisas relacionadas a essa área.

Referências

AMORIM, M. C. et al. Improving financial time series prediction using exogenous series and neural networks committees. In: INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS, 2010, Barcelona. *Proceedings*... Barcelona, 2010. p.1-8.

BENICASA, A. X.; PAIXAO, R. S. *Mineração de dados como ferramenta para descoberta de conhecimento*. Macapá: Faculdade Seama / Ministério Público do Estado do Amapá, 2006.

BOUMAN, R.; DONGEN, J. Van. *Pentaho solutions*: business intelligence and data warehousing with Pentaho and MySQL. Hoboken: Wiley, 2009.

CÔRTES S. C.; PORCARO R. M.; LIFSCHITZ S. *Mineração de dados*: funcionalidades, técnicas e abordagens. 2002. 35 f. Trabalho de Conclusão de Curso (Graduação) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2002.

FAYYAD, U.; PIATESKY-SHAPIRO, G.; SMYTH, P. *Advances in knowledge discovery and data mining*. California: AAAI/MIT Press, 1996.

FERNANDES, Elvis P.; COSTA JUNIOR, J. C. C.; CRISTIANO, V. *Fidelização de clientes usando inteligência artificial*. 2007. 139 f. Trabalho de Conclusão de Curso (Graduação) – Faculdades Associadas de São Paulo, 2007.

FRANK, A.; ASUNCION, A. *UCI Machine learning repository*. Irvine, CA: University of California, School of Information and Computer Science, 2010. Disponível em: [http://archive.ics.uci.edu/ml]. Acesso em: 30 set. 2011.

GOLDSCHMIDT, R.; PASSOS, E. *Data mining* : um guia prático. Rio de Janeiro: Campus, 2005.

HALL, M. et al. The WEKA data mining software: an update. *SIGKDD Explorations*, v. 11, p. 1-9, 2009.

HAN, J.; KAMBER, M. *Data mining concepts and techniques*. San Diego: Academic Press, 2001.

LENZ, Rafael A. *Utilizando técnicas de aprendizado de máquina para apoiar o teste de regressão*. 2009. 150 f. Dissertação (Mestrado em Informática)– Universidade Federal do Paraná, 2009.

LIMA, Lorena M. *Mineração de dados utilizando algoritmos genéticos adaptativos*.2009. 81 f. Trabalho de Conclusão de Curso (Graduação)-Universidade Federal da Bahia, 2009.

MATHWORKS. *MATLAB and simulink for technical computing*. Disponível em: <http://www.mathworks.com>. Acesso em: 30 set. 2011.

RAKOTOMALALA, R. *TANAGRA*: un logiciel gratuit pour l'enseignement et la recherche. In: ACTES DE EGC'2005, RNTI-E-3, 2005. v. 2, p. 697-702.