Análise de Técnicas de Mineração de Dados

Taylor Pablo Evaristo Silva[1], **Lívia Márcia Silva**[1]

[1]Universidade Presidente Antônio Carlos - Departamento de Ciência da Computação (UNIPAC)
Rua Palma Bageto Viol S/N – Barbacena – MG – Brasil

taylorpablo@hotmail.com, livimarcia@yahoo.com.br

Resumo. *O processo de descoberta de conhecimento em bases de dados (Knowledge Discovery in Databases KDD), incluindo fase de mineração de dados, vem sendo amplamente utilizado como ferramenta para auxiliar na tomada de decisão em áreas como crédito bancário e predições médicas. Neste trabalho este processo de KDD é estudado sendo como objetivo avaliar a utilização de um método de mineração de dados aplicado em uma base de dado.*

1. Introdução

Mineração de Dados é um ramo da computação que teve início nos anos 80, quando os profissionais das empresas e organizações começaram a se preocupar com os grandes volumes de dados informáticos estocados e inutilizados dentro da empresa. Nesta época, *Data Mining*(mineração de dados) consistia essencialmente em extrair informação de gigantescas bases de dados da maneira mais automatizada possível [Amo 2003].

Durante os últimos anos tem verificado um crescimento substancial da quantidade de dados armazenados em meios magnéticos. Segundo [Fayyad 1996] estes dados, produzidos e armazenados em larga escala, são inviáveis de serem lidos ou analisados por especialistas através de métodos tradicionais tais como planilhas de dados e relatórios informativos operacionais, onde o especialista testa sua hipótese contra a base de dados. Ou seja, as informações contidas nos dados não estão caracterizadas explicitamente, uma vez que sendo dados operacionais não interessam quando estudados individualmente. Logo, não bastava armazená-los, era preciso transformá-los em informações.

Estas informações tornaram-se essenciais para as empresas, já que as bases de dados deixaram de ser apenas repositórios de informações, passando a ser tratadas como patrimônio das mesmas.

Mineração de Dados é uma área de pesquisa multidisciplinar, incluindo tecnologia de bancos de dados, inteligência artificial, aprendizado de máquina, redes neurais, estatística, reconhecimento de padrões, sistemas baseados em conhecimento, recuperação da informação, computação de alto desempenho e visualização de dados.

Atualmente, *Data Mining* consiste sobretudo na análise dos dados após a extração, buscando-se por exemplo levantar as necessidades reais e hipotéticas de cada cliente para realizar campanhas de marketing. Assim, uma empresa de cartões de crédito, por exemplo, tem uma mina de ouro de informações: ela sabe os hábitos de compra de cada um dos seus seis milhões de clientes. O que costuma consumir,qual o seu padrão de gastos, grau de endividamento, etc. Para a empresa essas informações são extremamente úteis no estabelecimento do limite de crédito para cada cliente, e além disso, contém dados comportamentais de compra de altíssimo valor. Os seguintes pontos são algumas das razões por que o *Data Mining* vem se tornando necessário para uma boa gestão empresarial:

- Os volumes de dados são muito importantes para um tratamento utilizando somente técnicas clássicas de análise.
- O usuário final não é necessariamente um estatístico.
- a intensificação do tráfico de dados (navegação na Internet, catálogos *online*, etc) aumenta a possibilidade de acesso aos dados.

A mineração de dados falando simplesmente, trata-se de extrair ou minerar conhecimento de grandes volumes de dados. Muitas pessoas consideram o termo Mineração de Dados como sinônimo de *Knowledge Discovery in Databases* (KDD) ou Descoberta de Conhecimento em Banco de Dados. Na verdade, KDD é um processo mais amplo consistindo das seguintes etapas:

1. Limpeza dos dados: etapa onde são eliminados ruídos e dados inconsistentes.
2. Integração dos dados: etapa onde diferentes fontes de dados podem ser combinadas produzindo um único repositório de dados.
3. Seleção: etapa onde são selecionados os atributos que interessam ao usuário. Por exemplo, o usuário pode decidir que informações como endereço e telefone não são de relevantes para decidir se um cliente é um bom comprador ou não.
4. Transformação dos dados: etapa onde os dados são transformados num formato apropriado para aplicação de algoritmos de mineração (por exemplo, através de operações de agregação).
5. Mineração: etapa essencial do processo consistindo na aplicação de técnicas inteligentes afim de se extrair os padrões de interesse.

1.1. Motivação

A mineração de dados é extremamente importante, visto que permite a extração de conhecimento a partir de grandes volumes de dados. Dessa forma, pode-se perceber e reconhecer determinados padrões de clientes, por exemplo, e a empresa consegue conhecer melhor seus clientes. A mineração de dados pode ser usada em diversos tipos de empresas, como bancos, indústrias, comércios, entre outros.

1.2. Objetivos

Este trabalho tem por objetivo analisar o comportamento de uma base sobre uma das técnicas de mineração de dados. Objetiva também analisá-la e comparar seus resultados.

2. Estado da Arte

[Bala 1995] estudaram uma forma de aprendizado híbrido, usando algoritmo genérico e árvores de decisão para classificação. A ideia foi a integração do algoritmo AG GENESIS com população constantes e taxas de cruzamento e mutação respectivamente iguais a 0,6 e 0,001 com o algoritmo C4.5 para o procedimento de evolução. Os resultados experimentais foram apresentados para apresentar a eficácia da pesquisa em problemas complexos. Foram estudadas duas bases de dados. Uma delas composta de dados para reconhecimento de imagens faciais, apresentando erros de 27,5% e a outra reconhecimento visual de satélite, apresentando erro de 6,97%. Os resultados mostram bons desempenhos de classificação quando comparados com métodos clássicos para classificação que apresentam erros de 38,4% e 18, 5% respectivamente. [Lu 1995] publicaram um trabalho onde abordaram a aplicação de redes neurais para a classificação em mineração de

dados, dando ênfase as regras de extração. A base de dados trabalhada era composta de características de pessoas que seriam classificadas em grupos, tais como: idade, salário e possuir casa própria. Neste trabalho foi proposta a rede neural MPL (rede de múltiplas camadas compostas por neurônios do tipo perceptron) com algoritmo de retropropagação para aprendizado. Os resultados mostraram um erro menos utilizando estas redes quando comparadas ao algoritmo C4.5 de árvore de decisão. Porém, a rede neural precisou de um tempo maior de aprendizado. [Almeida e Dumonier 1996] publicaram um trabalho no qual apresentam uma abordagem estruturada de exploração de redes neurais, utilizando MPL, com algoritmo de aprendizado de retropropagação. O método foi utilizado para avaliação de riscos de inadimplência, avaliando 2412 empresas do setor de transporte de carga rodoviário francês. O desempenho foi comparado com o método de regressão logística(LOGIT). Foi concluído que o desempenho da rede neural implementada não foi significativamente superior ao desempenho do método estatístico, porém possui uma maior capacidade de generalização. [Fayyad 1996] publicou o trabalho *From Data Mining to Knowledge Discovery in databases* no qual descrevem como são relacionadas a mineração de dados e o KDD em um banco de dados, como em seus campos relacionados estatísticas e aprendizagem de máquina. Neste trabalho é conceituado que KDD é todo o processo de descoberta de conhecimento e a mineração de dados refere-se apenas uma fase deste processo. No trabalho são relatadas técnicas específicas para mineração de dados tais como árvore de decisão, regressão não linear e modelos de aprendizagem relacional. É discutido que não existe um método mais eficiente que sirva para todas as aplicações. A escolha do método vai variar de acordo com o objetivo da mineração de dados.

[Almeida e Dumonier 1996] e [SIQUEIRA] fazem uma comparação entre regressão logística, com o algoritmo LOGIT e redes neurais, aplicando o algoritmo de retropagação em uma rede MPL. As técnicas foram aplicadas a uma base de dados balanceada de 54 bancos brasileiros para a avaliação do risco de insolência. A técnica de rede neural não apresentou um fator diferencial que foi o de poder considerar a base de dados com campos vazios. A regressão logística necessita de base de dados com todos os campos não vazios.

[Zhang 1841] publicaram um trabalho propondo uma rede neural artificial para diagnóstico e detecção de falha em transformadores, considerando as concentrações de gases no óleo do transformador. Os dados são classificados de acordo com quatro diagnósticos. A rede neural utilizada foi a perceptron de múltiplas camadas (MPL) com o aprendizado feito pelo algoritmo de retropagação. As simulações foram feitas variando os parâmetros de entrada, o número de camadas escondidas e o número de nós de saída. A validação foi realizada com a técnica de validação cruzada. Os autores chegaram á conclusão de que quanto mais complexa a relação mais dados de treinamento são necessários e que aumentando a quantidade destes dados a Acurácia do modelo pode ser melhorada.

[Wang] publicaram um trabalho para diagnóstico de falhas em transformadores. Foi proposta uma classificação dos estados transformadores baseado em três formas:em sistemas especialistas, em redes neurais e em redes neurais conjugada com sistemas especialistas, chamadas autores de redes neurais especialistas. As simulações foram feitas com uma base de 210 dados. A rede neural utilizada foi a perceptron de múltiplas camadas(MPL) com o algoritmo de retropropagação para o treinamento. Os resultados do

trabalho mostram que o sistema conjugado tem melhor performance quando comparado com os resultados de classificação feita por cada sistema separadamente.

[Brammer 2011] e [BANZHAF] publicaram um trabalho onde apresentam uma comparação entre programação genética linear e a técnica de redes neurais, utilizando a rede MPL com o algoritmo de retropropagação resiliente para aprendizado, para mineração de dados médicos.o desempenho dos dois métodos foi compatível, sendo a programação genética linear considerada satisfatória na classificação e generalização dos dados.

3. Metodologia

Um dos maiores dilemas enfrentados por quaisquer sistemas de tomada de decisão é determinar um meio eficiente para produzir classificadores a partir de base de dados em relação ao tempo de processamento e à forma de representação simbólica simples e compreensível que facilite a análise do problema em questão.

Os classificadores baseados na árvore de decisão são um dos ramos na área de inteligência artificial. Mais especificamente, eles pertencem ao sub-campo de aprendizagem de máquina. Isto se deve à sua habilidade de aprender através de exemplos com o objetivo de classificar registros em uma base de dados.

3.1. Árvores de decisão

As árvores de decisão são amplamente utilizadas em algoritmos de classificação, e são representações simples do conhecimento, sendo um meio eficiente de construir classificadores que predizem ou revelam classes, ou informações úteis baseadas nos valores de atributos de um conjunto de dados. Eles são muito úteis em atividades de mineração de dados, isto é, o processo de extração de informações previamente desconhecida, a partir de grandes bases de dados. Aplicações desta técnica podem ser vista em diversas áreas, desde cenários de negócios até sistemas de piloto automático de aeronaves e diagnósticos médicos.[Pichiliani]

Uma árvore de decisão é essencialmente uma série de declarações if-elses, que quando aplicados a um registro de uma base de dados , resultam na classificação daquele registro. O mais interessante sobre o programa de árvores de decisão não é a sua construção a partir de classificação de um conjunto de treinamento, e sim a sua habilidade de aprendizado. Quando o treinamento é finalizado, é possível alimentar sua árvore de decisão construída a partir de exemplos com novos casos a fim de classificá-los [Curotto 2000].

Uma árvore de decisão é uma estrutura de árvore onde:

1. Cada nó interno é um atributo do banco de dados de amostras, diferente do atributo-classe.[Amo 2003]
2. As folhas são valores do atributo-classe.[Amo 2003]
3. Cada ramo ligando um nó- lho a um nó-pai é etiquetado com um valor do atributo contido no nó-pai. Existem tantos ramos quantos valores possíveis para este atributo.[Amo 2003]
4. Um atributo que aparece num nó não pode aparecer em seus nós descendentes.[Amo 2003]

3.2. Bases de Dados

A base de dados utilizada foi uma extração feita por Barry Becker a partir do banco de dados do Censo de 1994 dos EUA. A base é constituída por vários atributos que revelam as pessoas adultas dos Estados Unidos que tem uma renda menor ou maior que $ 50K/yr(50 mil dólares por ano). Os atributos da base de dados são: Idade, Classe de Trabalho, Educação Profissional, Estado Civil, Ocupação, Relação, Sexo, Raça, Ganho de Capital, Perda de Capital, Horas por Semana, Naturalidade, Probabilidade. A base está disponível no repositório do UCI (disponíveis em `http://www.ics.uci.edu/~mlearn/MLRepository.html`).

3.2.1. Ferramenta Utilizada

Para implementação dos métodos foi utilizado a ferramenta KNIME. KNIME (Konstanz Informação Miner) é um usuário-amigável e abrangente código-aberto de integração de dados, processamento, análise e plataforma de exploração. Desde o primeiro dia, KNIME foi desenvolvido utilizando práticas de engenharia de *software* rigorosos e é usado por profissionais da indústria e da academia em mais de 60 países.

4. Resultados

Os testes foram realizados com o programa knime, onde foram analisadas as bases de dados, as técnicas de datamining utilizada foram de arvores de decisão por ela aceitar dados categóricos. As propriedades alteradas no desenvolvimento dos teste foram:

- A coluna de classe tem que estar selecionada para a classe Probabilidade, da base de dados, onde os atributos têm que ser nominais que no caso será: menor e maior;
- A opção medida de qualidade esta selecionada para taxa de ganho;
- A opção Mínimo numero de registros por nó é a quantidade mínima de registros requeridos em cada nó para os testes;
- A opção Número para armazenar registros de vista é a quantidade selecionada para armazenar registros de visualizados;
- A opção *Number threads* pode explorar vários segmentos e processadores, assim, múltiplas ou núcleos. Isso pode melhorar o desempenho. O valor padrão é definido como o número de processadores ou núcleos disponíveis para KNIME. Se definido como 1, o algoritmo é feito seqüencialmente

Tabela 1. Primeiro teste

Opção	Quantidade
Mínimo numero de registros por nó	1
Número para armazenar registros de vista	50
Número de tópicos	1

Resultado: Acurácia = 77,603% , Erro=23,397%

Tabela 2. Segundo teste

Opção	Quantidade
Mínimo numero de registros por nó	3
Número para armazenar registros de vista	30
Número de tópicos	5

Resultado: Acurácia = 79,005% , Erro=20,995%

Tabela 3. Terceiro teste

Opção	Quantidade
Mínimo numero de registros por nó	30
Número para armazenar registros de vista	500
Número de tópicos	5

Resultado: Acurácia = 81,503% , Erro=18,497%

Tabela 4. Quarto teste

Opção	Quantidade
Mínimo numero de registros por nó	1
Número para armazenar registros de vista	500
Número de tópicos	100

Resultado: Acurácia = 75,934% , Erro=24,066%

Tabela 5. Quinto teste

Opção	Quantidade
Mínimo numero de registros por nó	900
Número para armazenar registros de vista	500
Número de tópicos	100

Resultado: Acurácia = 75,699% , Erro=24,301%

Tabela 6. Sexto teste

Opção	Quantidade
Mínimo numero de registros por nó	100
Número para armazenar registros de vista	500
Número de tópicos	900

Resultado: Acurácia = 81,666%, Erro=18,334%

Tabela 7. Sétimo teste

Opção	Quantidade
Mínimo numero de registros por nó	100
Número para armazenar registros de vista	100
Número de tópicos	100

`Resultado: Acurácia = 81,666% , Erro=18,664%`

Tabela 8. Oitavo teste

Opção	Quantidade
Mínimo numero de registros por nó	100
Número para armazenar registros de vista	100
Número de tópicos	1

`Resultado: Acurácia = 80,418% , Erro=19,582%`

Tabela 9. Nono teste

Opção	Quantidade
Mínimo numero de registros por nó	150
Número para armazenar registros de vista	999
Número de tópicos	15

`Resultado: Acurácia = 75,699% , Erro=24,301%`

A Figura 1 apresenta o gráfico contendo os resultados dos testes realizados, onde visualiza-se que a melhor acurácia obtida foi no sexto teste, e a pior foi no quinto teste.

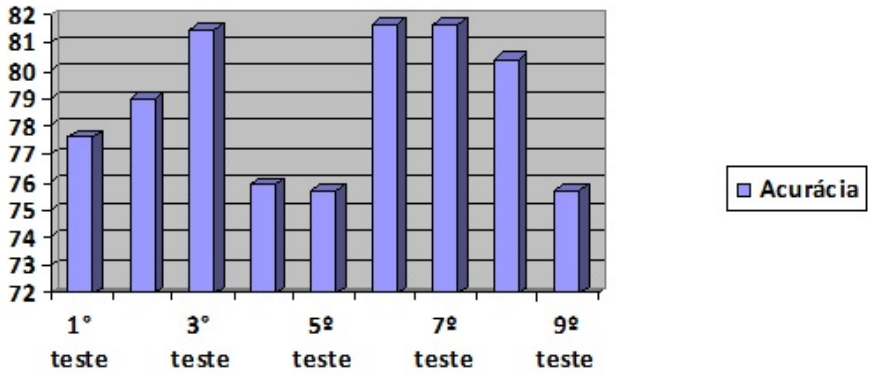

Figura 1. Resultados obtidos

Percebe-se que através dos testes obteve-se uma média em porcentagem de erros de 18,3 % a 24,3 %; houve uma média de acurácia de 75,6% a 81,6 %;quando aumentou-se em quantidade bem superior o numero de registros de vista em relação ao numero de registros por nó e sobre o numero de tópicos, a taxa de acurácia é melhor;quando aumenta-se o número de tópicos bem superior ao numero de registros por nó e quase o dobro em relação ao numero de registros de vista, a taxa de erro é bem menor; que quando iguala-se todas as opções, a taxa de acurácia é maior ;quando coloca-se o numero de tópicos bem

superior em relação ao numero de registros por nó, a taxa de erro é maior; que quando coloca-se o numero de registros por nó quase ao dobro em relação aos registros de vista e em bem quantidade elevada ao numero de tópicos, a taxa de erro é maior; que quando aumentamos elevadamente a opção numero de tópicos por vista em relação as outras duas opções, a taxa de erro é maior.

5. Conclusão

KDD é um processo de descoberta de conhecimento em bases de dados que tem como objetivo principal extrair conhecimento a partir de grandes bases de dados.

A etapa mais importante desse processo é a mineração de dados que se caracteriza pela existência de um algoritmo que diante da tarefa proposta será eficiente em extrair conhecimento implícito e útil de um banco de dados. Pode-se dizer que a mineração de dados é a fase que transforma dados puros em informações úteis.

Através dos teste conclui-se que quando iguala-se todas as opções, a taxa de acurácia é maior, e que quando aumentamos elevadamente a opção numero de tópicos por vista em relação as outras duas opções, a taxa de erro é maior.

Referências

R. AGRAWAL, R. SRIKANT, *Fast Algorithms for Mining Association Rules.Proc. 20th Int. Conf. Very Large Data Bases, VLDB, 1994*

R. AGRAWAL, R. SRIKANT, *Mining Sequential Patterns. In Proc. of 1995 Int.Conf. on Data Engineering, Taipei, Taiwan, March 1995.*

R. AGRAWAL, R. SRIKANT, *Mining Sequential Patterns : Generalizations and Performance Improvements. Proc. 5th EDBT, 3-17, 1996.*

R. SRIKANT, Q. VU, R. AGRAWAL, *Mining Association Rules with ItemConstraints Proc. of the 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, August 1997.*

FAYYAD, U., *From Data Mining to Know Ledge Discovery in Databases. American Association for Artificial Intelligence, 1996.*

BALA, J., *Hybrid Learning Using Genetic Algorithms and Decision Trees for Pattern Classification. Montreal, 1995. IJCAI Conference.*

LU, H., *NeuroRule: A Connectionist Approach to Data Mining, In: Proceedings of the 21st VLDB Conference, 1995.*

ALMEIDA, F.C. AND DUMONIER, P., *O uso de Redes Neurais em Avaliação de Riscos de Inadimplência, Revista de administração FEA/USP, São Paulo jan/mar 1996.*

ZHANG, Y., *Na Atificial Neural Network Approach to Transformer Fault Diagnosis. IEE Transactions on Power Delivery,1836-1841.*

WANG, Y., *An Artificial Neural Network Approach to Transformer Fault Diagnosis. IEE Transactions on Power Delivery. 1224-1229.*

BRAMMER, M., *A Comparision of Linear Genetic Programming an Neural and Neural Networks in Medical Data Mining,2001.*

CUROTTO, C.L., *Árvores de Decisão. Rio de Janeiro, 2000. Tese (Mestrado em engenharia Civil) Computação de Alto Desempenho e Sistemas Computacionais COPPE, Universidade Federal do Rio de Janeiro.*

PICHILIANI, MAURO, *Data Mining na Prática: Árvores de Decisão. Disponível em:* `http://imasters.com.br/artigo/5130/sql-server/data-mining-na-pratica-arvores-de-decisao`

S. DE AMO, *Curso de Data Mining, Programa de Mestrado em Ciência da Computação, Universidade Federal de Uberlândia, 2003. Disponível em:* `http://www.deamo.prof.ufu.br/CursoDM.html`