Parâmetros na escolha de técnicas e ferramentas de mineração de dados

2 authors, including:

Maria Madalena Dias
Universidade Estadual de Maringá
23 PUBLICATIONS **26** CITATIONS

Parâmetros na escolha de técnicas e ferramentas de mineração de dados

Maria Madalena Dias

Departamento de Informática, Universidade Estadual de Maringá, Av. Colombo, 5790, 87020-900, Maringá, Paraná, Brasil. e-mail: mmdias@din.uem.br

RESUMO. Apesar da existência de técnicas e ferramentas de mineração de dados, muitas organizações ainda desconhecem o quanto o computador pode dar suporte à tomada de decisão. A pouca utilização dessas técnicas e ferramentas pode estar relacionada à dificuldade na escolha da técnica e/ou ferramenta de mineração de dados mais adequada ao tipo de aplicação. A escolha da técnica de mineração de dados depende do problema de negócio a ser solucionado e das características dos dados disponíveis para análise, enquanto que na escolha da ferramenta de mineração de dados deve-se levar em consideração vários parâmetros, tais como: características gerais da ferramenta, conexão a bancos de dados, critérios de desempenho computacional, critérios de funcionalidade, critérios de usabilidade, etc. Neste artigo são apresentados parâmetros a serem considerados na escolha da técnica e da ferramenta de mineração de dados, sugeridos por vários autores. Também são mostrados os resultados obtidos com a aplicação de duas técnicas de mineração de dados.

Palavras-chave: técnicas de mineração de dados, ferramentas de mineração de dados.

ABSTRACT. Parameters of choice for data mining tools and techniques. Many companies do not realize the importance of computer back up for decision-making, despite the existence of data mining tools and techniques. The poor application of these techniques and tools can be related to the difficulty of choosing the most appropriate data mining tools and techniques for each application type. The choice of a data mining technique depends on the business problem to be solved and the data characteristics available for analysis, while the choice of a data mining tool, otherwise, depends on several parameters such as: tool general characteristics, data base connection, computational performance criteria, functionality criteria, using criteria, etc. In this paper, the parameters are established in the choice for data mining technique and tool, according to several authors' suggestions. The results achieved are also shown through the application of two data mining techniques.

Key words: data mining techniques, data mining tools.

Introdução

Durante várias décadas, desde a invenção do primeiro computador, o principal objetivo da utilização do computador é solucionar problemas operacionais da organização. A grande maioria das organizações ainda não possui meios de utilização dos recursos computacionais na tomada de decisão. Apesar da existência de grandes bancos de dados com muitas informações sobre o negócio da empresa, ainda são encontradas dificuldades na descoberta de conhecimento baseada nessas informações.

Essas dificuldades podem estar relacionadas a um dos seguintes fatores: falta de conhecimento da existência de técnicas de mineração de dados; complexidade na implementação e aplicação de uma técnica de mineração de dados; falta de ferramentas adequadas; alto custo das ferramentas de mineração de dados disponíveis no mercado; falta de parâmetros de referência na escolha da técnica e da ferramenta mais adequadas a cada problema a ser solucionado.

As técnicas de mineração de dados são aplicadas em sistemas de descoberta de conhecimento em banco de dados com o objetivo de extrair informações estratégicas escondidas em grandes bancos de dados, por meio da pesquisa dessas informações e da determinação de padrões, classificações e associações entre elas (Goebel e Gruenwald, 1999).

A seguir, nas próximas seções, são apresentados alguns conceitos de mineração de dados e descritas, sucintamente, tarefas e técnicas de mineração de dados; são discutidas algumas sugestões de como

escolher a técnica de mineração de dados mais adequada ao tipo de aplicação; são relacionadas características importantes na escolha de uma ferramenta de mineração de dados; é descrita uma metodologia para avaliação e seleção de software de mineração de dados; são relacionados parâmetros a serem considerados na escolha de uma técnica e de uma ferramenta de mineração de dados, respectivamente; são mostrados os resultados obtidos com a aplicação de duas técnicas de mineração de dados; e, finalmente, são apresentadas as considerações finais deste trabalho.

Mineração de dados

A mineração de dados pode ser considerada como uma parte do processo de Descoberta de Conhecimento em Banco de Dados (*KDD - Knowledge Discovery in Databases*). Segundo Goebel e Gruenwald (1999), o termo KDD é usado para representar o processo de tornar dados de baixo nível em conhecimento de alto nível, enquanto mineração de dados pode ser definida como a extração de padrões ou modelos de dados observados.

"Mineração de dados é a exploração e a análise, por meio automático ou semi-automático, de grandes quantidades de dados, a fim de descobrir padrões e regras significativos" (Berry e Linoff, 1997: 5). Os principais objetivos da mineração de dados são descobrir relacionamentos entre dados e fornecer subsídios para que possa ser feita uma previsão de tendências futuras baseadas no passado.

Os resultados obtidos com a mineração de dados podem ser usados no gerenciamento de informação, processamento de pedidos de informação, tomada de decisão, controle de processo e muitas outras aplicações.

A mineração de dados pode ser aplicada de duas formas: como um processo de verificação e como um processo de descoberta (Groth, 1998). No processo de verificação, o usuário sugere uma hipótese acerca da relação entre os dados e tenta prová-la aplicando técnicas como análises estatística e multidimensional sobre um banco de dados contendo informações passadas. No processo de descoberta não é feita nenhuma suposição antecipada. Esse processo usa técnicas, tais como: descoberta de regras de associação, árvores de decisão, algoritmos genéticos e redes neurais.

Tarefas desempenhadas por técnicas de mineração de dados

As técnicas de mineração de dados podem ser aplicadas a tarefas[1] como classificação, estimativa, associação, segmentação e sumarização. Essas tarefas são apresentadas de forma resumida na Tabela 1 (Dias, 2001).

Técnicas de mineração de dados

Harrison (1998) afirma que não há uma técnica que resolva todos os problemas de mineração de dados. Diferentes métodos servem para diferentes propósitos; cada método oferece suas vantagens e suas desvantagens. A familiaridade com as técnicas é necessária para facilitar a escolha de uma delas de acordo com os problemas apresentados. A Tabela 2 apresenta um resumo das técnicas de mineração de dados normalmente usadas.

Áreas de aplicação de técnicas de mineração de dados

A seguir, são relacionadas as principais áreas de interesse na utilização de mineração de dados (Viveros *et al*., 1996; Mannila, 1997; Cratochvil, 1999):

- Marketing. Técnicas de mineração de dados são aplicadas para descobrir preferências do consumidor e padrões de compra, com o objetivo de realizar marketing direto de produtos e ofertas promocionais, de acordo com o perfil do consumidor.
- Detecção de fraudes. Muitas fraudes óbvias (tais como, a compensação de cheque por pessoas falecidas) podem ser encontradas sem mineração de dados, mas padrões mais sutis de fraude podem ser difíceis de ser detectados, por exemplo, o desenvolvimento de modelos que predizem quem será um bom cliente ou aquele que poderá se tornar inadimplente em seus pagamentos.
- Medicina. Caracterizar comportamento de paciente para prever visitas, identificar terapias médicas de sucesso para diferentes doenças, buscar por padrões de novas doenças.
- Instituições governamentais. Descoberta de padrões para melhorar as coletas de taxas ou descobrir fraudes.
- Ciência. Técnicas de mineração de dados podem ajudar cientistas em suas pesquisas, por exemplo, encontrar padrões em estruturas moleculares, dados genéticos, mudanças globais de clima, oferecendo conclusões valiosas rapidamente.
- Controle de processos e controle de qualidade. Auxiliar no planejamento estratégico de linhas de produção e buscar por padrões de condições físicas na embalagem e armazenamento de produtos.

[1] Neste contexto, tarefa é um tipo de problema de descoberta de conhecimento a ser solucionado.

Tabela 1. Tarefas realizadas por técnicas de mineração de dados

Tarefa	Descrição	Exemplos
Classificação	Constrói um modelo de algum tipo que possa ser aplicado a dados não classificados a fim de categorizá-los em classes, o objetivo é descobrir um relacionamento entre um atributo meta (cujo valor será previsto) e um conjunto de atributos de previsão	Classificar pedidos de crédito Esclarecer pedidos de seguros fraudulentos Identificar a melhor forma de tratamento de um paciente
Estimativa (ou Regressão)	Usada para definir um valor para alguma variável contínua desconhecida	Estimar o número de filhos ou a renda total de uma família Estimar o valor em tempo de vida de um cliente Estimar a probabilidade de que um paciente morrerá baseando-se nos resultados de diagnósticos médicos Prever a demanda de um consumidor para um novo produto
Associação	Usada para determinar quais itens tendem a ser adquiridos juntos em uma mesma transação	Determinar que produtos costumam ser colocados juntos em um carrinho de supermercado
Segmentação (ou *Clustering*)	Processo de partição de uma população heterogênea em vários subgrupos ou grupos mais homogêneos	Agrupar clientes por região do país Agrupar clientes com comportamento de compra similar Agrupar seções de usuários Web para prever comportamento futuro de usuário
Sumarização	Envolve métodos para encontrar uma descrição compacta para um subconjunto de dados	Tabular o significado e desvios padrão para todos os itens de dados Derivar regras de síntese

Tabela 2. Técnicas de mineração de dados

Técnica	Descrição	Tarefas	Exemplos
Descoberta de Regras de Associação	Estabelece uma correlação estatística entre atributos de dados e conjuntos de dados	Associação	Apriori, AprioriTid, AprioriHybrid, AIS, SETM (Agrawal e Srikant, 1994) e DHP (Chen *et al.*., 1996).
Árvores de Decisão	Hierarquização dos dados, baseada em estágios de decisão (nós) e na separação de classes e subconjuntos	Classificação Regressão	CART, CHAID, C5.0, Quest (Two Crows, 1999); ID-3 (Chen *et al.*., 1996); SLIQ (Metha *et al.*., 1996); SPRINT (Shafer *et al.*., 1996).
Raciocínio Baseado em Casos ou MBR	Baseado no método do vizinho mais próximo, combina e compara atributos para estabelecer hierarquia de semelhança	Classificação Segmentação	BIRCH (Zhang *et al.*., 1996); CLARANS (Chen *et al.*., 1996); CLIQUE (Agrawal *et al.*., 1998).
Algoritmos Genéticos	Métodos gerais de busca e otimização, inspirados na Teoria da Evolução, onde a cada nova geração, soluções melhores têm mais chance de ter "descendentes"	Classificação Segmentação	Algoritmo Genético Simples (Goldberg, 1989); Genitor, CHC (Whitley, 1993); Algoritmo de Hillis (Hillis, 1997); GA-Nuggets (Freitas, 1999); GA-PVMINER (Araújo *et al.*., 1999).
Redes Neurais Artificiais	Modelos inspirados na fisiologia do cérebro, onde o conhecimento é fruto do mapa das conexões neuronais e dos pesos dessas conexões	Classificação Segmentação	Perceptron, Rede MLP, Redes de Kohonen, Rede Hopfield, Rede BAM, Redes ART, Rede IAC, Rede LVQ, Rede Counterpropagation, Rede RBF, Rede PNN, Rede Time Delay, Neocognitron, Rede BSB (Azevedo, 2000), (Braga *et al.*, 2000), (Haykin, 2001)

- Banco. Detectar padrões de uso de cartão de crédito fraudulento, identificar clientes "leais", determinar gastos com cartão de crédito por grupos de clientes, encontrar correlações escondidas entre diferentes indicadores financeiros.
- Apólice de seguro. Análise de reivindicações - determinar quais procedimentos médicos são reivindicados juntos, prever quais clientes comprarão novas apólices, identificar padrões de comportamento de clientes perigosos, identificar comportamento fraudulento.
- Transporte. Determinar as escalas de distribuição entre distribuidores, analisar padrões de carga.
- C e T (Ciência e Tecnologia). Avaliar grupos de pesquisa do país (Romão, 1999; Gonçalves, 2000; Dias, 2001).
- Web. Existem muitas pesquisas direcionadas à aplicação de mineração de dados na Web, tais como: (Loh *et al.*, 2000; Kosala e Blockeel, 2000; Ma *et al.*, 2000; Mobasher *et al.*, 2000; Sarawagi e Nagaralu, 2000; Spiliopoulou, 2000).

Como escolher a técnica de mineração de dados mais adequada

A escolha de uma técnica de mineração de dados a ser aplicada não é uma tarefa fácil. Segundo Harrison (1998), a escolha das técnicas de mineração de dados dependerá da tarefa específica a ser executada e dos dados disponíveis para análise. Berry e Linoff (1997) sugerem que a seleção das técnicas de mineração de dados deve ser dividida em dois passos:

1. Traduzir o problema de negócio a ser resolvido em séries de tarefas de mineração de dados;
2. Compreender a natureza dos dados disponíveis em termos de conteúdo e tipos de campos de dados e estrutura das relações entre os registros.

O primeiro passo na seleção da técnica de mineração de dados é, portanto, estabelecer uma meta comercial como, por exemplo, 'manter os clientes' e transformá-la em uma ou mais das tarefas de mineração de dados apresentadas anteriormente. Neste exemplo, a estratégia é identificar assinantes que tenham a intenção de desistir do serviço, descobrir suas razões para isso e fazer algum tipo de oferta especial que os agrade. Para o sucesso da estratégia, é preciso não somente identificar os assinantes que podem cancelar, mas dividi-los em grupos de acordo com seus motivos presumíveis para a desistência. A primeira tarefa é, obviamente, a classificação. Usando um conjunto de dados de treinamento com exemplos de clientes que cancelaram o serviço juntamente com exemplos daqueles que permaneceram, é possível construir um modelo capaz de rotular cada cliente como 'fiel' ou 'instável'.

O segundo passo, determinar as características dos dados em análise, tem como meta selecionar a técnica de mineração de dados que minimiza o número e dificuldades de transformação de dados para, a partir destes, obter bons resultados. A Tabela 3 mostra uma lista das características de dados baseada em (Berry e Linoff, 1997), que ajudará na escolha de uma abordagem de mineração de dados.

"Diferentes esquemas de classificação podem ser usados para categorizar métodos de mineração de dados sobre os tipos de bancos de dados a serem estudados, os tipos de conhecimento a serem descobertos e os tipos de técnicas a serem utilizadas" (Chen *et al.,* 1996, p.4), como pode ser visto a seguir:

- Com que tipos de bancos de dados trabalhar:
  Um sistema de descoberta de conhecimento pode ser classificado de acordo com os tipos de bancos de dados sobre os quais técnicas de mineração de dados são aplicadas, tais como: bancos de dados relacionais, bancos de dados de transação, orientados a objetos, dedutivos, espaciais, temporais, de multimídia, heterogêneos, ativos, de herança, banco de informação de Internet e bases textuais.
- Qual o tipo de conhecimento a ser explorado:
  Vários tipos de conhecimento podem ser descobertos por extração de dados, incluindo regras de associação, regras características, regras de classificação, regras discriminantes, grupamento, evolução e análise de desvio.
- Qual tipo de técnica a ser utilizada:
  A extração de dados pode ser categorizada de acordo com as técnicas de mineração de dados subordinadas. Por exemplo, extração dirigida a dados, extração dirigida a questionamento e extração de dados interativa. Pode ser categorizada, também, de acordo com a abordagem de mineração de dados subordinada, tal como: extração de dados baseada em generalização, baseada em padrões, baseada em teorias estatísticas ou matemáticas, abordagens integradas, etc.

A descoberta de regras de associação parece ser uma das técnicas de mineração de dados mais utilizada, sendo encontrada em diversas pesquisas (Agrawal e Srikant, 1994; Chen *et al.*, 1996; Holsheimer *et al.*, 1996; Viveros *et al.*, 1996; Mannila, 1997; Hipp *et al.*, 2000).

Tabela 3. Características de dados

Característica	Descrição	Técnicas de Mineração de Dados
Variáveis de categorias	São campos que apresentam valores de um conjunto de possibilidades limitado e predeterminado	Descoberta de regras de associação Árvores de decisão
Variáveis numéricas	São aquelas que podem ser somadas e ordenadas	Raciocínio baseado em casos (MBR) Árvores de Decisão
Muitos campos por registro	Este pode ser um fator de decisão da técnica correta para uma aplicação específica, uma vez que os métodos de mineração de dados variam na capacidade de processar grandes números de campos de entrada	Árvores de decisão
Variáveis dependentes múltiplas	Caso em que é desejado prever várias variáveis diferentes baseadas nos mesmos dados de entrada	Redes neurais
Registro de comprimento variável	Apresentam dificuldades na maioria das técnicas de mineração de dados, mas existem situações em que a transformação para registros de comprimento fixo não é desejada	Descoberta de regras de associação
Dados ordenados cronologicamente	Apresentam dificuldades para todas as técnicas e, geralmente, requerem aumento dos dados de teste com marcas ou avisos, variáveis de diferença etc.	Rede neural intervalar (*time-delay*) Descoberta de regras de associação
Texto sem formatação	A maioria das técnicas de mineração de dados é incapaz de manipular texto sem formatação	Raciocínio baseado em casos (MBR)

Ferramentas de mineração de dados

De acordo com Goebel e Gruenwald (1999), muitas ferramentas atualmente disponíveis são ferramentas genéricas da Inteligência Artificial ou da comunidade de estatística. Tais ferramentas geralmente operam separadamente da fonte de dados, requerendo uma quantidade significativa de tempo gasto com exportação e importação de dados, pré e pós-processamento e transformação de dados. Entretanto, segundo os autores, a conexão rígida entre a ferramenta de descoberta de conhecimento e a base de dados analisada, utilizando o suporte do SGBD (Sistema de Gerenciamento de Banco de Dados) existente, é claramente desejável. Para Goebel e Gruenwald (1999), as características a serem consideradas na escolha de uma ferramenta de descoberta de conhecimento devem ser as seguintes:

- A habilidade de acesso a uma variedade de fontes de dados, de forma *on-line* e *off-line*;
- A capacidade de incluir modelos de dados orientados a objetos ou modelos não padronizados (tal como multimídia, espacial ou temporal);
- A capacidade de processamento com relação ao número máximo de tabelas/tuplas/atributos;
- A capacidade de processamento com relação ao tamanho do banco de dados;
- Variedade de tipos de atributos que a ferramenta pode manipular; e
- Tipo de linguagem de consulta.

Goebel e Gruenwald (1999) propõem, também, um esquema de classificação de características que pode ser usado para estudar ferramentas de descoberta de conhecimento e de mineração de dados. Neste esquema, as características das ferramentas são classificadas em três grupos chamados características gerais, conectividade de banco de dados e características de mineração de dados. As Tabelas 4, 5 e 6 mostram como as características das ferramentas são classificadas de acordo com esses grupos.

Metodologia para avaliação e seleção de software de mineração de dados

Collier *et al.* (1999) apresentam uma estrutura para avaliar ferramentas de mineração de dados e descrevem uma metodologia para aplicação desta estrutura. As fases dessa metodologia estão representadas na Figura 1.

Tabela 4. Características gerais da ferramenta

Característica	Classificação
Produto	Nome e vendedor do produto de software
Status da Produção	P=Comercial, A=Alfa, B=Beta, R=Protótipo de Pesquisa
Status Legal	PD=Domínio Público, F=*Freeware*, S=*Shareware*
Licença Acadêmica	Se existe licença acadêmica livre disponível ou redução de custo
Demo	D=Versão Demo disponível para *download* na internet, R=Demo disponível através de requisição, U=Não-conhecido
Arquitetura	S=*Standalone*, C/S=Cliente/Servidor, P=Processamento Paralelo
Sistemas Operacionais	Lista de sistemas operacionais para os quais a versão atual do software pode ser obtida.

Tabela 5. Conectividade de bancos de dados da ferramenta

Característica	Classificação
Fontes de Dados	T=Arquivos texto Ascii, D=Arquivos Dbase, P=Arquivos Paradox, F=Arquivos Foxpro, Ix=Informix, O=Oracle, Sy=Sybase, Ig=Ingres, A=MS Access, OC=Conexão aberta de banco de dados (ODBC), SS=Servidor MS SQL, Ex=MS Excel, L=Lótus 1-2-3.
Conexão a BD	Onl=Online, Offl=Offline
Tamanho	S=Pequeno (até 10.000 registros), M=Mediano (10.000 a 1.000.000 registros), L=Grande (mais de 1.000.000)
Modelo	R=Relacional, O=Orientado a Objetos, 1= Uma Tabela
Atributos	Co=Contínuo, Ca=Categórico (valores numéricos discretos), S=Simbólico
Consulta	S=Linguagem de consulta estruturada (SQL ou derivada), Sp=Uma linguagem de consulta específica, G=Interface gráfica de usuário, N=Não aplicável, U=Não-conhecido

Tabela 6. Características de mineração de dados da ferramenta

Característica	Classificação
Tarefas Descobertas	Pré=Processamento de Dados (Amostragem, Filtragem), P=Predição, Regr=Regressão, Clã=Classificação, Clu=Agrupamento, A=Associação, Vis=Visualização do Modelo, EDA=Análise de Dados Exploratória
Metodologia de Descoberta	NN=Redes Neurais, GA=Algoritmos Genéticos, FS=Conjuntos Fuzzy, RS=Conjuntos Irregulares (Rough), St=Métodos Estatísticos, DT=Árvores de Decisão, RI=Indução de Regras, BN=Redes Bayseanas, CBR= Raciocínio Baseado em Casos
Interação Humana	A=Autônoma, G=Processo de descoberta guiado ao homem, H=Altamente interativo.

Quatro categorias de critérios para avaliar ferramentas de mineração de dados podem ser sugeridas: desempenho, funcionalidade, usabilidade e suporte de atividades principais de uma organização ou sistema (Collier *et al.*, 1999). A seguir, estas categorias são descritas.

- Desempenho: é a habilidade de manipular uma variedade de fontes de dados de maneira eficiente. A Tabela 7 relaciona critérios de desempenho computacional.
- Funcionalidade: é a inclusão de uma variedade de capacidades, técnicas e metodologias para mineração de dados. Ajuda avaliar o quanto a ferramenta adaptar-se-á a diferentes domínios de problema de mineração de dados. A Tabela 8 mostra critérios de funcionalidade.
- Usabilidade: é a acomodação de diferentes níveis e tipos de usuários sem perda de funcionalidade ou utilidade. Uma boa ferramenta deve fornecer parâmetros significativos para ajudar a depurar problemas e melhorar a saída. A Tabela 9 relaciona critérios de usabilidade.
- Suporte de atividades principais de uma organização ou sistema: esta categoria permite ao usuário desempenhar limpeza, manipulação, transformação, visualização de dados e outras tarefas para suporte à mineração de dados. A Tabela 10 mostra critérios para a categoria de suporte de atividades principais de uma organização ou sistema.

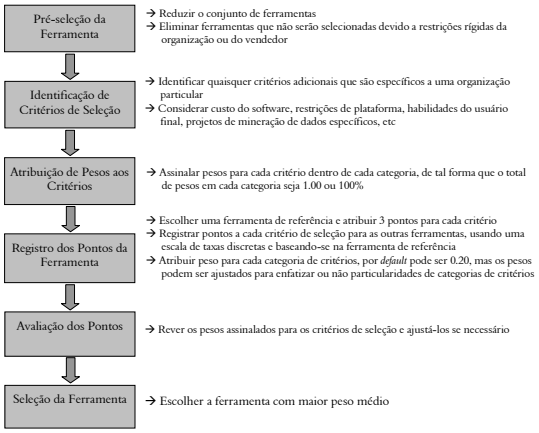

Figura 1. Fases da metodologia para seleção de ferramentas de mineração de dados

Tabela 7. Critérios de desempenho computacional

Critério	Descrição
Variedade de Plataforma	O software executa sobre uma grande variedade de plataformas computacionais? Ele executa sobre plataformas de usuário típicas de negócio?
Arquitetura de Software	O software usa arquitetura cliente-servidor ou arquitetura *stand-alone*? O usuário tem uma escolha de arquiteturas?
Acesso a Dados Heterogêneos	O software tem interface com uma variedade de fontes de dados (RDBMS, ODBC, CORBA etc)? Ele requer qualquer software auxiliar para fazer isto?
Tamanho dos Dados	O software escala para grandes conjuntos de dados? O desempenho é linear ou exponencial?
Eficiência	O software produz resultados em uma quantidade razoável de tempo relativo ao tamanho dos dados, as limitações do algoritmo e outras variáveis?
Interoperabilidade	A ferramenta facilita a interface com outras ferramentas de suporte KDD? Ela usa uma arquitetura padrão tal como CORBA ou alguma outra API proprietária?
Robustez	A ferramenta executa consistentemente sem parar? Se a ferramenta não realiza uma análise de mineração de dados, ela falha quando a análise parece estar quase completa?

Tabela 8. Critérios de funcionalidade

Critério	Descrição
Variedade Algorítmica	O software fornece uma variedade adequada de técnicas e algoritmos de mineração incluindo redes neurais, indução de regas, árvores de decisão, agrupamento, etc?
Metodologia Prescrita	O software ajuda o usuário a apresentar um estilo, metodologia de mineração passo a passo para ajudar a evitar resultados incorretos?
Validação do Modelo	A ferramenta suporta validação do modelo além da sua criação? A ferramenta encoraja validação como parte da metodologia?
Flexibilidade de Tipo de Dado	A implementação dos algoritmos manipula uma grande variedade de tipos de dados, dados contínuos sem amarração, etc?
Facilidade de Modificação dos algoritmos	O usuário tem a habilidade para modificar e sintonizar bem os algoritmos de modelagem?
Amostragem de Dados	A ferramenta permite amostragem randômica de dados para modelagem de predição?
Reportagem	Os resultados de uma análise de mineração resultam em uma variedade de caminhos? A ferramenta fornece resultados resumidos e resultados detalhados? A ferramenta seleciona registros de dados precisos que capacitam um perfil meta?
Exportação de Modelo	Após um modelo ser validado, a ferramenta fornece uma variedade de caminhos para ser exportada para uso em outro ambiente (por ex., programa C, SQL, etc.)?

Tabela 9. Critérios de usabilidade

Critério	Descrição
Interface do Usuário	A interface do usuário é fácil de usar para navegar e não complicada? A interface apresenta resultados de forma significativa?
Aprendizagem	A ferramenta é fácil de aprender? Ela é fácil de usar corretamente?
Tipos de Usuários	A ferramenta é projetada para usuários iniciantes, intermediários e avançados ou uma combinação de tipos de usuários? Ela é adequada para o tipo de usuário alvo? Ela é fácil de ser usada por analistas? Ela é fácil de ser usada por usuários finais?
Visualização de Dados	A ferramenta apresenta bem os dados? A ferramenta apresenta bem os resultados de modelagem? Existe uma variedade de métodos gráficos usados para comunicar informação?
Relatório de Erros	Os erros relacionados são significativos? As mensagens de erro ajudam o usuário na depuração dos problemas? A ferramenta acomoda bem os erros ou falsifica construção do modelo?
História de Ação	A ferramenta mantém uma história de ações realizadas no processo de mineração? O usuário pode modificar partes de sua história e re-executar o roteiro?
Variedade de Domínio	A ferramenta pode ser usada em uma variedade de indústrias diferentes para ajudar a solucionar uma variedade de tipos diferentes de problemas de negócio? A ferramenta foca bem sobre um domínio de problema? Ela foca bem sobre uma variedade de domínios?

Tabela 10. Critérios de suporte de atividades principais de uma organização ou sistema

Critério	Descrição
Limpeza de Dados	A ferramenta permite ao usuário modificar valores incorretos no conjunto de dados para desempenhar outras operações de limpeza de dados?
Substituição de Valores	A ferramenta permite substituição global de um valor de dado por outro (por ex., substituir 'M' ou 'F' por 1 ou 0)?
Filtragem de Dados	A ferramenta permite a seleção de subconjuntos dos dados baseando-se em critérios de seleção definidos pelo usuário?
Discretização	A ferramenta permite tornar dados contínuos em dados discretos para melhorar a eficiência da modelagem? A ferramenta requer que dados contínuos sejam discretizados ou esta decisão fica a critério do usuário?
Atributos de Derivação	A ferramenta permite a criação de atributos derivados baseando-se em atributos de herança? Existe uma grande variedade de métodos disponíveis para derivar atributos (por ex., funções estatísticas, funções matemáticas, funções booleanas, etc.)?
Definição de Dados Randômicos	A ferramenta permite definir dados randômicos antes da construção do modelo? A definição de dados randômicos é eficiente e efetiva?
Exclusão de Registros	A ferramenta permite exclusão de registros de entrada que podem estar incompletos ou podem predispor os resultados da modelagem de alguma forma? A ferramenta permite a exclusão de registros de segmentos de entrada da população? Se isto é possível, então a ferramenta permite que esses registros possam ser introduzidos facilmente mais tarde se necessário?
Manipulação de Espaços	A ferramenta permite manipular bem espaços? Ela permite que espaços sejam substituídos com uma variedade de valores derivados? Ela permite que espaços sejam substituídos com um valor definido pelo usuário? Se isto é possível, pode ser globalmente bem como valor por valor?
Manipulação de Metadados	A ferramenta apresenta ao usuário descrições, tipos e códigos categóricos de dados, fórmula para derivar atributos, etc.? Se isto é possível, a ferramenta permite que o usuário manipule esse metadados?
Realimentação de Resultados	A ferramenta permite que os resultados de uma análise de mineração sejam retornados para uma outra análise na construção de mais modelos?

Parâmetros na escolha de técnicas de mineração de dados

Basicamente, os principais parâmetros a serem considerados na escolha de uma técnica de mineração de dados são:

- Tipo de problema de descoberta de conhecimento a ser solucionado: este parâmetro é obtido com a definição da tarefa de mineração de dados, que deve estar de acordo com os objetivos definidos para a descoberta de conhecimento em questão.
- Características dos dados: a adequação da técnica de mineração de dados às características dos dados (ver Tabela 3) visa, principalmente, minimizar as dificuldades geralmente encontradas na transformação de dados.
- Forma de aplicação da mineração de dados: a mineração de dados pode ser aplicada como um processo de verificação, onde o usuário tenta provar uma hipótese acerca da relação entre os dados, ou como um processo de descoberta, onde não é feita nenhuma suposição antecipada. Existem técnicas mais propícias para o processo de verificação (análises estatística e multidimensional) e outras para o processo de descoberta (regras de associação, árvores de decisão, algoritmos genéticos e redes neurais). No entanto, pesquisas atuais mostram a aplicação de algoritmo genético no processo de verificação (Romão, 2002).
- Disponibilidade de ferramenta de mineração de dados: um problema de descoberta de conhecimento pode ser solucionado, em determinados casos, com a aplicação de mais de um tipo de técnica de mineração de dados. Assim, pode ser escolhida uma técnica ou outra dependendo da ferramenta disponível.

Parâmetros na escolha de ferramenta de mineração de dados

Parâmetros para a escolha de ferramenta de mineração de dados, sugeridos por Goebel e Gruenwald (1999) e Collier *et al.* (1999), foram descritos anteriormente.

Todos os parâmetros sugeridos são bastante relevantes, no entanto, deve-se considerar a real necessidade da organização e feita uma análise cuidadosa de custo e benefício na aquisição e utilização de uma ferramenta desse tipo.

Outro parâmetro muito importante a ser considerado é o suporte técnico da empresa fornecedora. Geralmente, uma ferramenta de mineração de dados, por mais "amigável" que seja a sua interface, apresenta dificuldades em sua utilização, inerentes ao tipo de aplicação. O usuário de uma ferramenta de mineração de dados precisa ter um bom conhecimento sobre a área de negócio da organização, sobre as técnicas de mineração de dados implementadas pela ferramenta e de todo o processo de descoberta de conhecimento. Além disso tudo, o usuário precisa saber analisar os resultados obtidos pela ferramenta e onde e como utilizá-los.

Os parâmetros, sugeridos pelos autores citados e neste trabalho, podem ser resumidos como segue:

- Habilidade de acesso a uma variedade de fontes de dados, de forma *on-line* e *off-line*;
- Capacidade de incluir modelos de dados orientados a objetos ou modelos não padronizados;
- Capacidade de processamento com relação ao tamanho do banco de dados e ao número máximo de tabelas/tuplas/atributos;
- Variedade de tipos de atributos que a ferramenta pode manipular;
- Tipo de linguagem de consulta;
- Capacidade de acomodação de diferentes níveis e tipos de usuários sem perda de funcionalidade ou utilidade;
- Capacidade de adaptar-se a diferentes domínios de problema de mineração de dados;
- Capacidade de desempenhar limpeza, manipulação, transformação, visualização de dados e outras tarefas para suporte à mineração de dados;
- Custo x benefício;
- Suporte técnico da empresa fornecedora.

Estudo de caso

O estudo de caso realizado teve como principal objetivo a utilização de parâmetros relacionados anteriormente na escolha da técnica de mineração de dados.

A escolha da técnica de mineração de dados iniciou-se com a definição de objetivos da descoberta de conhecimento a ser realizada. Em seguida, verificou-se que o problema a ser solucionado poderia ser tratado tanto como uma tarefa de associação quanto de classificação.

As técnicas de mineração de dados foram aplicadas sobre dados de Programas de Pós-Graduação do Brasil contidos em bases de dados da Capes do ano de 1998.

O objetivo da descoberta de conhecimento definido foi o estudo da relação entre fomento, nível de formação do quadro funcional do Programa de Pós-Graduação e a produtividade média dos pesquisadores.

Considerando que os dados são classificados como variáveis de categorias, por terem sido discretizados, Descoberta de Regras de Associação e Árvore de Decisão são as técnicas mais indicadas para a solução das tarefas identificadas.

Foram utilizados três algoritmos que implementam as técnicas de mineração de dados escolhidas. O primeiro foi o "apriori", que implementa a técnica de Descoberta de Regras de Associação, realizando assim a tarefa de associação. Este algoritmo foi implementado por Gonçalves (2000) e aplicado através do protótipo do ambiente ADesC (Dias, 2001).

O segundo algoritmo foi o C4.5 (Quinlan, 1993), que implementa a técnica de Árvore de Decisão e realiza a tarefa de classificação. Este algoritmo foi estudado e aplicado no Trabalho de Final de Curso de Igarashi (2002), utilizando o programa C4.5 Decision Tree Generator (C4.5DTG - Versão 1.00), obtido através do endereço www.sff.sdf.br.

O terceiro algoritmo aplicado foi o J48.PART, que também implementa a técnica de Árvore de Decisão e realiza a tarefa de classificação. Este algoritmo foi estudado e aplicado no Trabalho de Graduação de Centeio (2002), utilizando o Weka 3, que é um software de aprendizagem de máquina em Java (Weka, 2001).

As ferramentas foram escolhidas considerando a confiabilidade dos algoritmos implementados e o fato das mesmas serem de domínio público. Portanto, não foi feita nenhuma avaliação das ferramentas em relação aos parâmetros definidos anteriormente.

A seguir são apresentados os resultados obtidos com a aplicação das técnicas de Descoberta de Regras de Associação e de Árvore de Decisão.

1) Descoberta de regras de associação

Além dos dados de entrada contidos na base de dados da Capes, já preparados e formatados de acordo com a exigência do algoritmo *¨apriori¨* utilizado, os parâmetros suporte mínimo e grau de confiança são necessários. Os valores mínimos de suporte e de confiança utilizados nos experimentos foram 8% e 40%, respectivamente.

A Tabela 11 relaciona as regras geradas para este estudo (Dias, 2001).

No estudo pode-se observar que o aumento relativo na quantidade de bolsas de um Programa implica aumento na Produção, mas há um nível de saturação para esta regra (quando a média de bolsa por aluno está entre 25 e 30 meses). É interessante observar que esta constatação vai de encontro à política de redução do tempo máximo de bolsa adotada pela Capes.

As regras referentes à Titulação do corpo Docente e Discente não permitem o mesmo tipo de análise, dado que apresentaram relações absolutas e independentes.

2) Árvore de decisão

A técnica de Árvore de Decisão foi aplicada através dos algoritmos C4.5 e J48.PART do Weka, tendo como parâmetros de entrada os dados dos Programas de Pós-Graduação do Brasil, os mesmos utilizados no algoritmo "apriori". A Tabela 12 mostra os resultados obtidos com a aplicação desses algoritmos.

O estudo mostra que quase 70% dos pesquisadores produzem em média de 2 a 3 publicações por ano, independentemente de sua formação e de sua área de atuação. Sendo que para as áreas de Ciências Exatas e da Terra, Engenharias, Ciências Biológicas e Ciências Agrárias, este percentual é um pouco maior. As regras geradas mostram também que a produção anual média por pesquisador é maior nas áreas de Lingüística, Letras e Artes e Ciências Humanas.

A partir das regras geradas, podemos concluir que o atributo "grau de formação dos pesquisadores" não pode ser considerado como atributo de previsão do atributo meta. Outro atributo que foi sugerido como um possível atributo previsor, "média de meses de bolsa por aluno", nem sequer foi incluído nas regras geradas pelo algoritmo de classificação.

Considerações finais

A mineração de dados surgiu com o objetivo principal de dar suporte à tomada de decisões na empresa. Portanto, a aplicação de técnicas de mineração de dados em sistemas de descoberta de conhecimento em banco de dados busca a descoberta de regras e padrões em dados que trarão o conhecimento suficiente e adequado para aquelas pessoas responsáveis pela tomada de decisões na empresa.

Tabela 11. Regras geradas com a aplicação da técnica descoberta de regras de associação

Regra	Suporte	Confiança
Se o Programa possui de 21 a 40 mestres, então ele produz de 2 a 3 publicações em média por pesquisador por ano	21,73%	45,18%
Se o Programa possui de 19 a 24 meses de financiamento por aluno concluído, então ele produz de 2 a 3 publicações em média por pesquisador	19,62%	53,76%
Se o Programa possui de 25 a 30 meses de financiamento por aluno concluído, então ele produz de 2 a 3 publicações por pesquisador	16,88%	48,19%
Se o Programa possui de 25 a 30 meses de financiamento por aluno concluído, então ele possui de 21 a 40 Mestres	15,19%	43,37%

Tabela 12. Regras geradas com a aplicação da técnica árvore de decisão

Regra	Confiança
Se a grande área for Ciências Exatas e da Terra, então a produtividade do Programa é de 2 a 3 publicações em média por pesquisador	81,5%
Se a grande área for Engenharias, então a produtividade do Programa é de 2 a 3 publicações em média por pesquisador	76,1%
Se a grande área for Ciências Biológicas, então a produtividade do Programa é de 2 a 3 publicações em média por pesquisador	74,2%
Se a grande área for Ciências Agrárias, então a produtividade do Programa é de 2 a 3 publicações em média por pesquisador	72,6%
Se a grande área for Ciências Sociais Aplicadas, então a produtividade do Programa é de 2 a 3 publicações em média por pesquisador	67,7%
Se a grande área for Ciências Humanas e o total de mestres for de 41 a 70, então a produtividade do Programa é de 2 a 3 publicações em média por pesquisador	64%
Se a grande área for Lingüística, Letras e Artes, então a produtividade do Programa é de 4 a 6 publicações em média por pesquisador	74,1%
Se a grande área for Ciências Humanas, então a produtividade do Programa é de 4 a 6 publicações em média por pesquisador	56,5%
A produtividade do Programa é de 2 a 3 publicações em média por pesquisador, independentemente da área e da formação dos pesquisadores	67,3%

O usuário de um sistema de descoberta de conhecimento em banco de dados precisa ter um sólido entendimento do negócio da empresa para ser capaz de selecionar corretamente os subconjuntos de dados e as classes de padrões mais interessantes.

Os resultados obtidos com a aplicação dos algoritmos de associação e classificação mostram que um pesquisador produz em média 2 a 3 publicações por ano, independentemente de sua formação e da média de meses de bolsa por aluno de mestrado do programa. Os resultados dos algoritmos de classificação mostram ainda que a área de atuação do pesquisador pode ser considerada como um atributo previsor da média de publicação por pesquisador por ano.

Através do estudo de caso realizado, pode-se concluir que, para o tipo de aplicação estudado, Ciência e Tecnologia, as técnicas de mineração de dados Descoberta de Regras de Associação e Árvore de Decisão podem ser utilizadas com sucesso.

A tarefa de classificação deve ser usada quando o objetivo é descobrir atributo(s) previsor(es) para um ou poucos atributos meta. A vantagem da tarefa de associação é a correlação que é feita entre todos os atributos, podendo resultar na descoberta de atributos previsores para mais de um atributo meta.

Este artigo apresentou uma visão geral das áreas de mineração de dados e descoberta de conhecimento e relacionou alguns parâmetros a serem considerados na escolha de técnicas e de ferramentas de mineração de dados, baseados em sugestões de diversos autores. Mostrou, também, os resultados da aplicação de duas técnicas de mineração sobre dados dos cursos de Pós-Graduação do Brasil do ano de 1998.

Referências

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules. *In:* INTERNATIONAL CONFERENCE ON VERY LARGE DATABASES, 20., 1994, Santiago do Chile. *Proceedings…* Santiago do Chile, 1994. p. 487-499.

AGRAWAL, R. *et al.*. Automatic subspace clustering of high dimensional data for data mining applications. *In*: ACM SIGMOD CONFERENCE, Seattle, 1998. *Proceedigns…* Seattle: ACM Press, 1998. p. 94-105.

ARAÚJO, D.L.A. *et al.* A parallel genetic algorithm for rule Discovery in large databases. *IEEE Systems, Man and Cybernetics Conference*, Tokyo. v. 3, p. 940-945, 1999.

AZEVEDO, F.M. *et al. Redes neurais com aplicações em controle e em sistemas especialistas*. Florianópolis: Visual Books, 2000.

BERRY, M.J.A.; LINOFF, G. *Data mining techniques*. New York: John Wiley & Sons, Inc. 1997.

BRAGA, A.P. *et al. Redes neurais artificiais: teoria e aplicações*. Rio de Janeiro: Livros Técnicos e Científicos Editora S.A., 2000.

CENTEIO, S.J.D.M. *Mineração de dados usando WEKA*. Monografia (Graduação) - Curso de Bacharelado em Ciência da Computação da Universidade Estadual de Maringá, Maringá, 2002.

CHEN, M.S. *et al.* Data mining: an overview from database perspective. *IEEE Transactions on Knowledge and Data Engineering,*, v. 8, n.6, p. 866-883, 1996.

COLLIER, K. *et al.* A methodology for evaluating and selecting data mining software. *In*: HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 32., 1999, Hawaii. *Proceedigns…* Hawaii: University of Hawaii, 1999.

CRATOCHVIL, A. *Data mining techniques in supporting decision making*. Master Thesis - Universiteit Leiden, Leiden, 1999.

DIAS, M.M. *Um modelo de formalização do processo de desenvolvimento de sistemas de descoberta de conhecimento em banco de dados*. Tese (Doutorado) - Curso de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2001.

FREITAS, A.A. A genetic algorithm for generalized rule induction. In: WSC3, ON-LINE WORLD CONFERENCE ON SOFT COMPUTING, HOSTED ON THE INTERNET, 3., 1998, Cranfield. *Proceedings…* Cranfield: Cranfield University, 1999, p. 340-353.

GOEBEL, M.; GRUENWALD, L. A survey of data mining and knowledge discovery software tools. *ACM SIGKDD*, San Diego, v. 1, n. 1, p. 20-33, 1999.

GOLDBERG, D.A. *Genetic algorithms in search, otimization, and machine learning*. New Jersey: Addison-Wesley, 1989.

GONÇALVES, A.L. *Utilização de técnicas de mineração de dados na análise dos grupos de pesquisa no Brasil*. Dissertação (Mestrado) - Curso de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina, Florianópolis, 2000.

GROTH, R. *Data mining*. Englewood Cliffs: Prentice Hall, Inc., 1998.

HARRISON, T.H. *Intranet data warehouse*. São Paulo: Editora Berkeley Brasil, 1998.

HAYKIN, S. *Redes neurais:* princípios e prática. Porto Alegre: Bookman, 2001.

HILLIS, D.B. Using a genetic algorithm for multi-hypothesis tracking. *In:* INTERNATIONAL CONFERENCE ON TOOLS WITH ARTIFICIAL INTELLIGENCE (ICTAI'97), 9., 1997, Newport Beach. *Proceedings…* Newport Beach, 1997.

HIPP, J. *et al.* Algorithms for association rule mining - a general survey and comparison. *ACM SIGKDD*, Boston, v 2, n. 1, p. 58-64, 2000.

HOLSHEIMER, M. *et al.* A perspective on databases and data mining. *In*: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING,1., 1996, Menlo Park. *Proceedings...* Menlo Park: AAAI Press, 1996, p. 447-467.

IGARASHI, W. *Análise e aplicação de uma técnica de mineração de dados*. Monografia (Graduação) - Curso de Bacharelado em Informática da Universidade Estadual de Maringá, Maringá, 2002.

KOSALA, R.; BLOCKEEL, H. Web mining research: a survey. *ACM SIGKDD*, Boston, v. 2, n. 1, p. 1-15, 2000.

LOH, S. *et al.* Concept-based knowledge Discovery in texts extracted from the web. *ACM SIGKDD*, Boston, v. 2, n. 1, p. 29-39, 2000.

MA, Y. *et al.* Web for data mining: organizing and interpreting the discovered rules using the web. *ACM SIGKDD*, Boston, v. 1, n 1, p.16-23, 2000.

MANNILA, H. Methods and problems in data mining. *In:* INTERNATIONAL CONFERENCE ON DATABASE THEORY, 6., Delphi, 1997. *Proceedings…* Delphi: Lecture Notes in Computer Science, 1997. p. 41-55.

MEHTA, M. *et al*. SLIQ: a fast scalable classifier for data mining. *In:* EDBT'96, INTERNATIONAL CONFERENCE ON EXTENDING DATABASE TECHNOLOGY, 5., 1996, Avignon, 1996, *Proceedings…* Avignon: Lecture Notes in Computer Science, 1996, p. 18-32.

MOBASHER, B. *et al.* Automatic personalization based on web usage mining. *Communications of the ACM*, New York, v. 43, n. 8, p. 142-151, 2000.

QUINLAN, Ross. C4.5: *Programs for Machine Learning*. San Francisco: Morgan Kaufmann, 1993.

ROMÃO, W. *et al*. Extração de regras de associação em C e T: o algoritmo Apriori. *In:* ENCONTRO NACIONAL EM ENGENHARIA DE PRODUÇAO, 19., ICIE - INTERNATIONAL CONGRESS OF INDUSTRIAL ENGINEERING, 5., Rio de Janeiro. *Proceedings...* Rio de Janeiro, 1999.

ROMÃO, W. *Descoberta de conhecimento relevante em banco de dados sobre ciência e tecnologia*. Tese (Doutorado) - Curso de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina, Florianópolis, 2002.

SARAWAGI, A.; NAGARALU, S.H. Data mining models as services on the internet. *ACM SIGKDD*, Boston, v. 2, n. 1, p. 24-28, 2000.

SHAFER, J.C. *et al*. SPRINT: a scalable parallel classifier for data mining . *In:* VLDB CONFERENCE, 1996, Bombay. *Proceedings…* Bombay: IIT Bombay, 1996, p. 544-555.

SPILIOPOULOU, M. Web usage mining for web site evaluation. *Communications of the ACM*, New York, v. 43, n. 8, p.127-134, 2000.

TWO CROWS CORPORATION. Introduction to data mining and knowledge discovery. *Technical Report*. 3 ed., 1999.

VIVEROS, M.S. *et al.* Applying data mining techniques to a health insurance information system. *In*: VLDB CONFERENCE, 22., 1996, Bombay. *Proceedings…* Bombay: IIT Bombay, 1996. p. 286-295.

WEKA. *Waikato environment for knowledge analysis*, Disponível em: <http://www.cs.waikato.ac.nz/ml/weka>. Acesso em nov. 2001.

WHITLEY, D. A Genetic algorithm tutorial. *Technical Report*, 1993.

ZHANG, T. *et al*. BIRCH: an efficient data clustering method for very large database. *In*: ACM SIGMOD CONFERENCE, 1996, Montreal. *Proceedings…* Montreal: Le Centre Sheraton, 1996. p. 103-114.

Received on October 10, 2002.
Accepted on November 22, 2002.