DATA MINING: ALGUMAS QUESTÕES EPISTEMOLÓGICAS

Paulo Sérgio de Souza Coelho
Faculdades Ibmec – RJ
Avenida Rio Branco, 108 – 2º andar. Rio de Janeiro – RJ
psergio@ibmecrj.br

Resumo

Data Mining é uma área do conhecimento relativamente nova que trata do problema da extração de informações (conhecimento) a partir de repositórios digitais de dados. O conceito de KDD é menos conhecido mas é mais amplo e engloba o processo que vai desde a amostragem dos dados (seleção) até a interpretação do conhecimento expresso pelos modelos estimados. Estas áreas (KDD e Data Mining) são geralmente tratadas como áreas da ciência da computação, mas este artigo apresenta uma fundamentação teórica obtida a partir do cruzamento de opiniões especializadas onde outras ciências são consideradas. Para descrever os problemas que podem ser resolvidos a partir de suas técnicas, foi criada uma taxonomia. Diversos exemplos bem sucedidos do uso de Data Mining são citados e utilizados para apoiar a hipótese de que há rigor metodológico nos seus procedimentos. Finalmente, fazem-se alguns comentários sobre as críticas provenientes da econometria.
Palavras-Chaves: Data Mining; Taxonomia; Metodologia.

Abstract

Data Mining is a new knowledge area witch deal information (knowledge) extraction from digital data repository. The KDD concept is less known but more global, considering the whole process from data selection (sampling) until estimated model validation and interpretation. These areas (KDD and Data Mining) appear like computing science issues, but in this paper is presented a theoretical foundation obtained from several sources, including other sciences. A Taxonomy was created to better describe the problems relative these areas. A lot of well succeeded examples are related and used to support the hypothesis there is methodological rigor in Data Mining procedures. Finally, some comments are made about the econometrics criticizes.
Keywords: Data Mining; Taxonomy; Methodology.

1. INTRODUÇÃO

O problema fundamental relativo a Data Mining é normalmente referido como extração de conhecimento a partir de dados. Esta é uma área nova do conhecimento que surge a partir da interação de ciências principalmente quantitativas, como computação e estatística, mas também cognitivas como psicologia.

Muitas das que são consideradas suas técnicas já eram conhecidas antes do seu surgimento, como por exemplo, Regressão Linear ou Redes Neurais. Estas técnicas podem ser consideradas Data Mining porque tratam do problema que esta área do conhecimento humano se propõe a trabalhar. E estas técnicas são revistas segundo alguns princípios que norteiam a área de Data Mining, como o problema da abundância de dados ou a interatividade do processo em busca do modelo mais adequado.

O uso de Data Mining como uma metodologia de pesquisa tem sido discriminado, especialmente no Brasil. Muitos teóricos, estatísticos e econométricos, têm criticado o uso desta ferramenta para construção de conhecimento científico. Entretanto, diversas aplicações bem sucedidas das técnicas de Data Mining como ferramentas de gestão dão subsídios para garantir que os procedimentos de descoberta de conhecimento são efetivos. Neste contexto, por que não utilizar Data Mining como uma ferramenta para pesquisa?

As duas próximas seções apresentam os conceitos e definições relativas a este

ambiente. A Seção 4 apresenta como as técnicas de Data Mining podem ser utilizadas para problemas específicos, onde foi descrita uma taxonomia para estes problemas. Este artigo traz ainda exemplos do uso de técnicas de Data Mining como ferramenta de gestão na Seção 5, e uma discussão sobre princípios básicos para uma pesquisa que utilize como metodologia estas técnicas na Seção 6, onde algumas considerações disponíveis na literatura econométrica são utilizadas para tentar garantir que não há perda de rigor metodológico com estes procedimentos.

2. CONCEITOS E DEFINIÇÕES

Um conceito adequado para Data Mining deve necessariamente envolver os termos conhecimento e dado. Neste sentido, os dados representam a matéria prima e o conhecimento é o produto final da indústria do Data Mining. Os dados são fatos disponíveis que, ou estão em uma forma estruturada e num ambiente digital, como bancos de dados operacionais, ou podem ser assim codificados. O conhecimento descoberto pelo processo de Data Mining é uma generalização que pode ser obtida através dos dados. Esta generalização é também comumente chamada de padrão, no sentido que é uma estrutura recorrente nos dados sendo analisados. Além disso, dado que o fenômeno gerador dos fatos é ignorado ou apenas parcialmente conhecido, o produto obtido, conhecimento, é oculto. Por isso, usa-se também o termo descoberta de conhecimento para estas atividades.

Vamos examinar três conceitos, obtidos a partir da literatura especializada sobre o tema:

> Simploriamente falando, Data Mining refere-se à extração ou "mineração" de conhecimento a partir de grandes quantidades de dados. (HAN e KAMBER, 2001, p. 5, tradução nossa)
>
> Data Mining é definido como o processo de descoberta de padrões nos dados. O processo precisa ser automático ou (mais usualmente) semi-automático. Os padrões descobertos devem ser significativos pois devem trazer alguma vantagem, geralmente no sentido econômico. Os dados são invariavelmente apresentados em quantidades substanciais. (WITTEN e FRANK, 2005, p. 5, tradução nossa)
>
> Data Mining é um processo iterativo no qual o progresso é definido pela descoberta, através de métodos manuais e/ou automáticos. (WESTPHAL e BLAXTON, 1998, p. 6, tradução nossa)

Os dois primeiros conceitos enfatizam a questão da abundância dos dados disponíveis, que é uma preocupação particular dentro da área. A abundância de dados é frequentemente tratada como o problema da escalabilidade, onde a escala reflete diretamente o volume dos dados. Data Mining, enquanto área do conhecimento, tem como um dos objetos de estudo a eficiência das técnicas, mais especificamente dos algoritmos, diante de grandes volumes de dados. Neste sentido, diversas técnicas clássicas de análise de dados foram recriadas ou adaptadas de maneira a tornarem-se eficientes no aspecto da escalabilidade, ou seja, novos algoritmos foram desenvolvidos de maneira que métodos clássicos pudessem ser aplicados em grandes bancos de dados eficientemente.

Os dois últimos conceitos enfatizam a questão da iteração do processo de Data Mining, no sentido que a análise é constante e o processo é desenhado em diversos passos. Este processo total é preferivelmente chamado de **KDD – *Knowledge Discovery in Databases,*** ou Descoberta de Conhecimento em Bancos de Dados. Este processo é dividido em etapas que vão desde a localização e extração dos dados até a compreensão do conhecimento modelado. A construção do modelo de conhecimento, que é como preferimos definir Data Mining, é apenas uma das etapas intermediárias (BERRY e LINOFF, 2004, BERSON, SMITH e THEARLING, 1999, HAN e KAMBER, 2001 e WEISS e INDURKHYA, 1998). Esta diferenciação é apenas uma questão semântica. De fato, o analista de Data Mining está diretamente envolvido com todo o processo, motivo pelo qual poderia ser considerado analista de KDD. Assim, não se faz necessário, do ponto de vista prático, esta diferenciação. Entretanto, estamos considerando-a aqui para estruturar melhor os conceitos.

2.1. KDD

A diferenciação dos conceitos de KDD e de Data Mining, aqui considerada, não é

unânime na literatura. Para alguns autores, estes conceitos podem ser considerados os mesmos. Por exemplo, Witten e Frank (2005) nem chegam a estabelecer o conceito de KDD, e Westphal e Blaxton (1998) fazem uma rápida menção ao termo (p. 5) sem maiores diferenciações. Giudici (2003) reconhece o termo KDD, inclusive com uma visão de desenvolvimento histórico, concordando com a definição que foi feita aqui. O autor chega até a considerar outras referências nesta linha (p. 2), mas fala de Data Mining no restante do livro como sinônimo de KDD: "Data Mining não é apenas o uso de um algoritmo computacional ou uma técnica estatística; é um processo de *Business Intelligence* que pode ser usado juntamente com o que é fornecido pela tecnologia da informação para dar apoio às decisões da companhia" (p. 3, tradução nossa). O conceito de *Business Intelligence* é explorado na Seção 3.

Dentro do quadro que especifica o conceito de KDD, pode-se considerar que este representa uma metodologia de trabalho de Data Mining, ou seja, um conjunto de passos pré-definidos que devem ser seguidos para se obter o resultado final: conhecimento significativo e interpretável. Dentro desta metodologia, Data Mining é apenas uma das fases do processo completo que, ao todo, compreende mais três etapas de trabalho: Seleção, Preparação e Transformação e, finalmente, depois da fase de modelagem, Interpretação (HAN e KAMBER, 2001) – sendo as duas primeiras etapas relativas a dados e a última relativa ao conhecimento modelado pelo Data Mining.

A Figura 1 apresenta um esquema com as etapas contidas no processo de KDD. As quatro etapas estão indicadas acima dos ícones representativos dos produtos de cada etapa, que são ao todo cinco. Os nomes destes produtos aparecem abaixo dos respectivos ícones. O produto inicial de cada etapa equivale ao produto final da etapa imediatamente anterior. Há apenas duas exceções a esta consideração. Na primeira etapa, que não tem etapa anterior, o produto inicial é a fonte dos dados, ou seja, o ambiente observado. A outra exceção é a última etapa, que não tem etapa posterior e, portanto, seu produto final é o produto final de todo o KDD, o conhecimento.

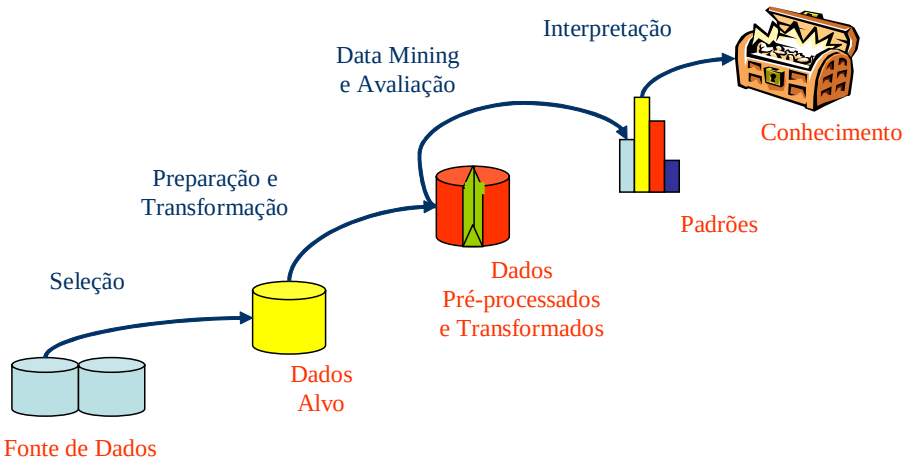

Figura 1: Etapas do KDD. Adaptado de Han e Kamber (2001, p. 6)

A primeira etapa refere-se à seleção dos dados provenientes de algum(as) fonte(s) de dados. Uma fonte de dados não é necessariamente um Sistema de Banco de Dados ou um *Data Warehouse*. É possível extrair conhecimento a partir de conjuntos de dados que estejam em outros repositórios, como planilhas, tabelas construídas a partir de questionários, pesquisa na internet, etc. De qualquer maneira, é necessário coletar os dados que são considerados relevantes e eliminar aqueles que forem considerados desnecessários. A eliminação pode parecer desnecessária, como se a etapa de Data Mining fosse capaz de avaliar os dados fazendo eventuais eliminações, o que não é sempre verdade. Algumas técnicas específicas de Data Mining são capazes de fazer tais eliminações, mas outras não, que podem até ter seu

rendimento prejudicado em função da redundância. Por isso, dados que de antemão sabe-se que não são relevantes ao problema que está sendo analisado devem ser evitados. Dependendo de características dos algoritmos de análise que serão utilizados, a redundância de dados pode atrapalhar a análise, como o problema da multicolinearidade para a regressão (um problema bastante comum, discutido em diversos textos de estatística ou econometria). Além disso, a redução no volume é sempre importante porque muitas vezes a análise da totalidade dos dados pode ser proibitiva em termos de tempo e de capacidade computacional disponíveis. Para contextualizar, quando a fonte de dados é um *Data Warehouse* é comum descartar até 90% do volume inicial dos dados, já que a redundância nestes sistemas (e a amplitude das informações) é, por definição, muito grande.

A segunda etapa do processo de descoberta do conhecimento é possivelmente a etapa mais importante dentre todas. O pré-processamento, como também é chamada esta etapa, é fundamental para o passo seguinte, de aquisição do conhecimento (Data Mining). Um pré-processamento mal realizado reduz, podendo até eliminar, as chances de uma modelagem eficiente (PYLE, 2001). Nesta etapa os dados são preparados para a modelagem, resolvendo-se problemas como os de redundância (que não tenha sido percebido na etapa anterior), inconsistência e ausência de valores. Por este motivo, esta etapa é também chamada de limpeza. Percebe-se então que quando estes problemas não são resolvidos, ou se são resolvidos de forma inconsistente ou incompleta, a etapa de modelagem de dados poderá apresentar resultados insatisfatórios, contraditórios ou inválidos. Não existem procedimentos prontos para esta etapa, requerendo do analista a sensibilidade de buscar possíveis problemas e encontrar soluções. Além disso, na maioria dos casos não há nenhuma maneira de estimar a qualidade do resultado, recaindo novamente sobre a sensibilidade do analista avaliações subjetivas para estas estimativas. Por exemplo, quando valores desconhecidos estão sendo substituídos é comum que não se tenha como estimar a precisão do valor substituído, ou caso se tenha alguma estimativa, seja ela imprecisa. Pyle (2001) aponta ainda que esta etapa pode ser a mais demorada de todo o processo de KDD.

A outra parte desta etapa, a transformação dos dados, é responsável pela garantia de que haverá compatibilidade com a ferramenta de modelagem a ser utilizada. Diferentes técnicas ou diferentes ferramentas computacionais podem ser empregadas na etapa de modelagem, e o formato de entrada dos dados depende das restrições do sistema a ser utilizado. Se for necessário modelar os dados utilizando diferentes sistemas que estabeleçam restrições específicas para o formato dos dados, será necessário realizar diversas transformações, que tomarão como entrada o resultado do pré-processamento (único), gerando diferentes saídas, uma para cada uma das restrições impostas.

A etapa de Data Mining é responsável pela captação e estruturação das características e padrões dos dados que estão sendo analisados na forma de algum modelo matemático-estatístico. Estas características e padrões podem, dentro de limites estimados através de intervalos de confiança e testes de hipóteses, ser utilizadas para projeções de fatos não observados. Esta etapa realiza, além da construção dos modelos, as avaliações deste modelo através de critérios pré-estabelecidos como precisão ou robustez. Podemos considerar que esta etapa está efetivamente relacionada com a descoberta do conhecimento. As técnicas de Data Mining podem ser classificadas segundo sua funcionalidade fundamental: descrição ou predição.

Finalmente, a última etapa, Interpretação, é responsável pela adequação da saída da ferramenta de modelagem às necessidades do usuário. Algumas ferramentas de visualização e de navegação nos modelos obtidos podem ser utilizadas. Esta etapa permite ao analista compreender o conhecimento modelado e validar, de modo empírico, o modelo obtido. Em geral, o analista usa a expertise no ambiente modelado, no mínimo o senso comum, para fazer esta validação.

Observe que estas atividades têm uma ordem bem estabelecida. Entretanto, em alguma etapa podem ser evidenciados problemas ocorridos em alguma etapa anterior. Assim,

o processo de KDD pode incluir a repetição de alguma(s) etapa(s). Um exemplo típico é a etapa de preparação de dados. É possível que algum problema nos dados só venha a ser percebido na fase de interpretação do modelo gerado depois da fase de Data Mining. A preparação de dados teria de ser refeita novamente, assim como todas as etapas subseqüentes.

Uma outra alteração comum no fluxo de KDD é quando alguma etapa pode ser suprimida como, por exemplo, a etapa de transformação dos dados. Existem algumas ferramentas que fazem todo o processo de KDD, desde a extração dos dados até a interpretação do modelo. São as chamadas suítes (de Data Mining ou de KDD). Nestes casos não é preciso fazer a transformação dos dados, pois a própria ferramenta o fez no momento da seleção ou extração.

2.2. FUNDAMENTAÇÃO

Voltando aos conceitos de Data Mining citados no início da seção, Witten e Frank (2005) tratam Data Mining como uma extensão direta de uma área da Inteligência Computacional chamada Aprendizado de Máquina. Entretanto, parece mais adequado aceitar o Aprendizado de Máquina como parte de uma das disciplinas que compõem a área de Data Mining, como pode ser visto na Figura 2. Aprendizado de Máquina é uma das linhas de ação da Inteligência Artificial, que juntamente com Banco de Dados e Visualização são as disciplinas da área de Computação que formam a base de Data Mining. Esta extensão é importante, porque outras linhas da Inteligência Artificial também estão relacionadas, como Representação do Conhecimento ou Agentes Inteligentes (BITTENCOURT, 2001).

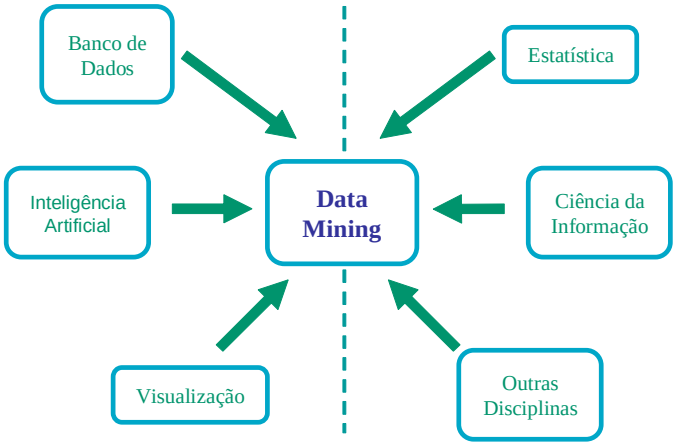

Figura 2: Data Mining e disciplinas correlatas. Adaptado de Han e Kamber (2001, p. 9)

Por outro lado, Estatística e Ciência da Informação são disciplinas também fundamentais, mas que estão mais voltadas para a formalização dos procedimentos de indução e de avaliação de modelos. Esta diferenciação em termos de computação e estatística é feita por Giudici (2003), que apresenta as disciplinas que formam uma base para Data Mining nestes dois grupos principais, separados pela linha vertical na Figura 2: do lado esquerdo as disciplinas de computação e do lado direito as disciplinas de estatística.

O que está sendo chamado de “Outras Disciplinas” na Figura 2 refere-se a eventuais contribuições vindas da Economia, Biologia ou Psicologia (HAN e KAMBER, 2001).

3. CONTEXTUALIZAÇÃO HISTÓRICA

Cientistas históricos apontam que o desenvolvimento da civilização humana se dá em função da posse da comunicação (HABERMAS, 2002). A comunicação permite que os aprendizados e as conquistas de uma geração sejam transmitidos para as gerações seguintes. Assim, a civilização avança, em função do avanço das ciências individuais, pois as descobertas e invenções formam uma base de conhecimentos que é utilizada para a projeção dos novos aprendizados e conquistas, que irão incorporar o conjunto do conhecimento. Este é o processo cíclico de desenvolvimento da humanidade, completamente dependente da

comunicação das experiências e seus resultados, sem estes positivos ou negativos (experiências bem ou mal sucedidas, respectivamente).

A maior parte dos dados gerados atualmente pelas atividades realizadas nos mais diversos setores da ação humana está sendo armazenada em dispositivos eletrônicos, geralmente relacionados a computadores. O nível de informatização (uso de sistemas eletrônicos de informação) e, consequentemente, o volume de dados armazenados tem aumentado de maneira sistemática.

No caso das organizações, todas as atividades são geradoras de diversos tipos de dados que podem ser armazenados, em função das estratégias técnicas dos sistemas de informação que os manipula. Por exemplo, os dados de uma organização geralmente incluem cadastros de produtos/serviços, clientes, fornecedores e parceiros, que muito provavelmente estarão armazenados, pois as operações mais básicas dependem destes dados. Num outro nível, as transações que são realizadas entre a organização e os clientes, fornecedores e/ou parceiros também pode ser armazenadas: vendas, pedidos de entrega ou de fornecimento, alterações no estoque, etc. Outro tipo de dados que também é muito importante é relativo ao nível dos registros financeiros e contábeis em geral.

O avanço da tecnologia de informação acelerou a dinâmica das organizações, de maneira que o volume de atividades que são realizadas cresce constantemente. Além disso, as operações tornam-se mais complexas, de maneira que mais detalhes tornam-se importante. O volume de dados tem aumentado não só pela quantidade de registros, mas também pela variedade de tipos de dados que precisam (e podem) ser armazenados. A questão da complexidade dos tipos de dados que estão sendo armazenados também tem sido uma forte restrição que a tecnologia de informação tenta resolver. Documentos em forma de desenhos, figuras, imagens, áudio e até vídeos, por exemplo, também estão sendo armazenados e requerem tratamento especial. Métodos específicos se fazem necessários não só para o armazenamento e consulta, mas, sobretudo, para a análise (objeto fundamental de Data Mining).

O armazenamento dos dados é feito originalmente para registro dos acontecimentos e para futuras **consultas** (*queries*). Essas consultas aos bancos de dados são realizadas ao nível de tabelas e registros específicos. Por exemplo, os registros de uma determinada tabela que atendem a uma determinada condição, ou algumas estatísticas como média ou totalização dos valores de um determinado atributo (variável). Para realizar consultas como estas foram desenvolvidas linguagens específicas, das quais a mais utilizada é conhecida como SQL – *Structured Query Language* (DATE, 1999). Consultas atendem ao nível operacional, pois trazem informações atômicas, como o cadastro de um cliente ou o histórico logístico de um determinado pedido.

Os dados podem ou não ser registrados ao longo do tempo. Um sistema de informações envolvido diretamente com as atividades operacionais da organização costuma armazenar apenas os dados específicos destas operações, como por exemplo, o estoque de determinado produto e os pedidos deste produto que estão sendo processados. O procedimento operacional pode ser processar cada um dos pedidos, um por vez. Quando um pedido é processado o estoque é atualizado. Do ponto de vista operacional é bastante armazenar as informações atualizadas, no caso, o estoque final, e retirar o produto processado da fila de pedidos que ainda o serão. Entretanto, do ponto de vista gerencial pode ser fundamental saber o histórico dos pedidos que já foram processados – para análise de informações como tempo de processamento, qualidade, etc. E, do ponto de vista estratégico, é importante guardar informações de como o estoque varia ao longo do tempo, e o cruzamento de variáveis como produto e cliente ou vendedor e região. Estas informações podem permitir análises que indiquem possíveis falhas operacionais ou simplesmente a percepção de que determinadas modificações podem aumentar a eficiência dos processos da organização. Assim, do ponto de vista operacional, basta um sistema que registre os valores dos dados atuais. Do ponto de vista estratégico/gerencial pode ser importante guardar estas mesmas

informações num contexto temporal. Kimball (1996) já apontava estas questões e apresentava uma solução que veio a se tornar um recurso básico para a organização moderna, que são os ***Data Warehouses***.

A partir de então, o desenvolvimento da tecnologia de armazenamento e acesso de dados diferencia os sistemas de **banco de dados operacionais** (dinâmico) e os *Data Warehouses*, ou armazéns de dados (estáticos). Os *Data Warehouses* tornam-se importantes (imprescindíveis para grandes corporações) para as atividades de análise, sobretudo sobre a ótica da otimização dos recursos computacionais envolvidos. Os bancos de dados operacionais (relacionais) são projetados para facilitar as atividades de armazenamento, consulta e alteração (atualização) de informações. O volume e a freqüência que estas operações são realizadas não permitem operações de análise nos bancos operacionais. Por outro lado, os *Data Warehouses* possuem uma estrutura de armazenamento de dados que privilegia a análise, sem preocupações operacionais de evitar a duplicidade (KIMBALL, 1996 e THOMSEM, 2002).

Os *Data Warehouses* oferecem as mesmas ferramentas de consulta que o sistema de banco de dados operacional e uma facilidade a mais: os dados podem ficar estruturados em forma de um cubo que é uma estrutura que facilita a observação dos dados. Esta estrutura é chamada de OLAP – *On Line Analytical Process*, ou cubo OLAP. O acesso à informação na forma de cubos mostra-se muito eficiente sob o aspecto de versatilidade e velocidade, se comparado ao uso de tabelas (planilhas) de informações ou resultados de consultas a bancos de dados (THOMSEM, 2002).

As ferramentas OLAP são úteis para análises de segmentos que podem ser representados através das dimensões do cubo (tipicamente centros de custo, localidades, departamentos, produtos e tempo). Entretanto estas ferramentas não são capazes de fazer a extração de todo o conhecimento contido nos *Data Warehouses*. Uma das dificuldades destas ferramentas é a dependência com a forma que os dados tenham sido estruturados no sistema, pois as análises estão limitadas a operações de agrupamento de valores em função das dimensões do cubo. Apesar destas limitações, as estruturas OLAP dos *Data Warehouses* são adequadas para as atividades de modelagem do conhecimento.

Diante do crescimento do volume dos repositórios (bancos de dados operacionais, *Data Warehouses* e outros), torna-se necessário realizar análises mais profundas das informações armazenadas. Convém distinguir os conceitos de consulta e análise, pois este último deve ser considerado como uma metodologia que permitirá não apenas a recuperação de informações (que é a consulta), mas o estabelecimento de modelos que representarão o que chamaremos de **conhecimento**. Observe que é a análise dos dados que permite realizar uma projeção de valores desconhecidos, ainda que limitada e passiva de erros. Por exemplo, a partir de uma análise de uma série histórica de dados é possível fazer previsões de fatos (índices, taxas, valores em geral) que ainda não são conhecidos ou que ainda não aconteceram (e, dependendo da metodologia de análise empregada, esta previsão pode vir acompanhada de todo um aparato inferencial para definir margens de erro e índices de qualidade). O objetivo destas análises é obter o conhecimento contido numa base de dados. Esta modelagem está relacionada a medidas de similaridade, co-ocorrência (correlação), relações de dependência, de ordem, etc.

Devido ao grande volume de dados que é armazenado nestes repositórios, a descoberta de conhecimento não pode ser conseguida a partir de simples consultas ao banco de dados. Torna-se necessário o suo, e eventual desenvolvimento, de modelos computacionais e estatísticos específicos. Neste contexto é que surge Data Mining como uma área específica do conhecimento, conforme descrita anteriormente.

As ferramentas de Data Mining modelam (não apenas agrupam ou segmentam) as informações a partir das massas de dados, vindas de “qualquer” origem e de “qualquer” forma, gerando conhecimento. Este conhecimento pode ser expresso através de um conjunto de regras ou estruturas gráficas como grafos, redes ou árvores. O tipo de conhecimento e a

maneira como é expresso depende da técnica e da metodologia de análise empregada.

Todas estas tecnologias de armazenamento e análise de dados têm sido enquadradas em um grupo de ferramentas computacionais que está sendo chamado de **BI – *Business Intelligence***, ou Inteligência de Negócios. A interface ao usuário que é oferecida por estes sistemas costuma ser muito avançada, permitindo uma poderosa manipulação dos dados a partir de uma interação com o sistema bastante simplificada, sem exigir conhecimentos avançados de tecnologia (MILLER et al, 2002 e SERRA, 2002).

Giudici (2003) relata esta contextualização histórica de maneira mais objetiva, tratando da evolução da área de BI, conforme pode ser visto na Figura 3. A seqüência desenhada representa não só o avanço em termos de evolução histórica, mas também em termos de capacidade de informação e dificuldade de implementação, estas duas últimas características mais importantes do ponto de vista do usuário destes sistemas. Neste sentido, o autor aponta para o *trade-off* entre estas características: os sistemas mais capazes (baseados em Data Mining) são os mais difíceis de serem implementados.

Figura 3: Evolução das Ferramentas de BI. Adaptado de Giudici (2003, p. 4)

4. TAXONOMIA PARA DATA MINING

Na literatura especializada em Data Mining existe uma grande quantidade de algoritmos, técnicas, métodos e heurísticas que são descritas de forma torrencial. É muito difícil, principalmente para um leitor iniciante, contextualizar todos os termos novos dentro de uma estrutura de contexto. Para que poder descrever como Data Mining funciona, desenvolveu-se uma taxonomia que permitisse isolar estes termos em grupos dentro de uma hierarquia de conceitos.

Um dos maiores problemas que se encontra na busca de uma taxonomia como esta é dar nome aos níveis taxonômicos. Em função disto, definimos que os **algoritmos** de Data Mining estão classificados no contexto de uma determinada **técnica**. A técnica reúne diversos algoritmos que tenham em comum algumas heurísticas ou estratégias de atuação. Os **problemas** representam um nível acima das técnicas, pois representam um tipo de padrão que se deseja obter. Finalmente, existem as **funcionalidades**, nas quais os problemas estão inseridos. Na Figura 4 é possível observar as duas funcionalidades de Data Mining: Descritiva e Preditiva, assim como as principais técnicas: Regras de Associação, Cluster, Classificação e Previsão. Até este nível não há sobreposição, e então a estrutura hierárquica da Figura 4 não apresenta ciclos. A tentativa de descrever as técnicas mais importantes tornaria esta estrutura mais confusa, pois existem técnicas que estão relacionadas com mais de um problema. Por exemplo, a técnica de Redes Neurais pode ser utilizada para problemas de Classificação e Previsão (na funcionalidade Preditiva) ou ainda, de Cluster (na funcionalidade Descritiva).

As funcionalidades de Data Mining estão de acordo com Han e Kamber (2001). Esta concepção é bem semelhante à divisão clássica da Estatística em termos de Descritiva e Inferencial. Assim, a Funcionalidade Descritiva está relacionada com atividades de caracterização dos dados que estão sendo analisados, enquanto a Funcionalidade Preditiva está relacionada com a generalização (indução) das observações presentes nos dados observados para criação de modelos que possam ser usados em dados não observados para realizar predições.

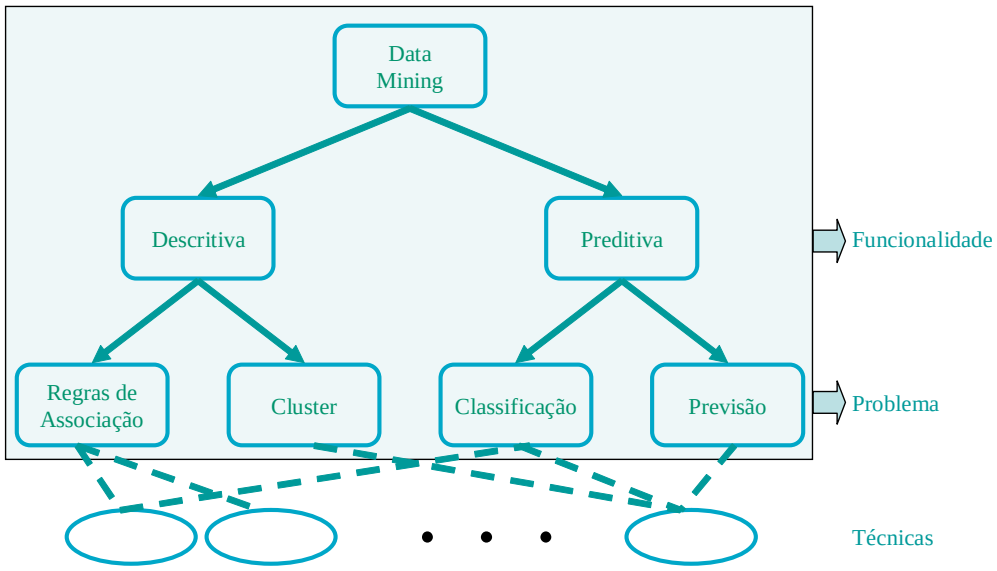

Figura 4: Taxonomia para técnicas de Data Mining

Os problemas estão mais diretamente relacionados com o tipo de modelo que se deseja estabelecer. Enquanto os dois problemas da funcionalidade descritiva são bem diferentes entre si, e sempre descritos de maneira totalmente independente, os da funcionalidade preditiva não são tão diferenciados pela literatura. Em geral, as técnicas preditivas são tratadas como sendo de um único problema, chamado de "Classificação e Previsão", como por exemplo, em Han e Kamber (2001). Isto ocorre fundamentalmente por que os principais algoritmos de Previsão podem ser usados para Classificação, e vice-versa. Um exemplo disso é a técnica conhecida como KNN – *K – Nearest Neighbor* (K vizinhos mais próximos), que é faz a predição para um determinado caso usando uma quantidade determinada (*k*) de exemplos que lhe estejam mais próximos.

As técnicas não estão descritas na Figura 4 por um motivo além do problema de não haver, a partir deste nível, hierarquia entre técnicas e problemas. A quantidade de técnicas de Data Mining é muito grande, e uma descrição completa destas técnicas não é nosso objetivo.

4.1. REGRAS DE ASSOCIAÇÃO

O problema de obter Regras de Associação foi proposto e solucionado inicialmente por Agrawal, Imielinksy e Swami (1993), para realizar o que ficou conhecido como Análise de Cesta de Mercado (*Market Basket Analysis*). A questão era identificar um padrão de comportamento de compra analisando apenas o histórico dos tíquetes de venda (lista de itens vendidos em cada transação). Este padrão tem a forma de uma regra: SE <condição> ENTÃO <conseqüência>, onde <condição> e <conseqüência> são conjuntos de itens de compra (*itemset*). Uma regra como esta estabelece que se o comprador adquirir o conjunto de itens expresso em <condição> ele adquirirá o conjunto de itens expressos em <conseqüência>. Algumas medidas de qualidade da regra são fornecidas como, por exemplo, o suporte – percentual de casos observados em que a regra é verificada.

A solução apresentada pelos autores foi o algoritmo chamado de Apriori. Este algoritmo é um marco importante dentro da Ciência da Computação, porque definiu uma heurística importante, conhecida como Propriedade Apriori (HAN e KAMBER, 2001, p. 231, 426). O algoritmo ainda é bastante utilizado até hoje, apresentando algumas vantagens sobre algoritmos desenvolvidos posteriormente.

A técnica proposta originalmente foi estendida para outras situações em que a análise de co-ocorrências poderia ser aplicada, extrapolando portanto a análise de cestas de mercado. Existe inclusive uma extensão para Análise de Correlação (HAN e KAMBER, 2001, p. 225).

4.2. CLUSTER

Não existe ainda uma tradução universalmente aceita para o termo *cluster* no sentido do problema de Data Mining. Uma tradução imediata é segmentação, entretanto, deve-se estar atento para evitar confusão com o sentido aplicado em Marketing, pois neste contexto segmentação está mais relacionado com o método de Classificação, pois está se lidando com uma variável que determina os segmentos, como a classe econômica ou a região geográfica. Agrupamento seria outra tradução possível, e é usado por alguns autores (por exemplo, Goldschmidt e Passos, 2005). Este termo é preferido.

O problema do agrupamento é separar as observações da base de dados que está sendo analisada e grupos (*clusters*). Cada grupo é uma coleção de elementos (observações) que são similares entre si, e bem diferentes dos elementos dos outros grupos. Similaridade e diferença (que são opostas mas podem ser definidas nos mesmos termos) são descrições para pares de elementos. As similaridades e diferenças entre os grupos são obtidas a partir da acumulação das medidas entre todos os elementos dos grupos. Estas medidas são geralmente relatadas através de distância numérica, o que pode requerer um tratamento prévio sobre as observações (fase do KDD de pré-processamento) para que esta distância possa ser calculada (transformação dos atributos categóricos).

Han e Kamber (2001, p. 335-393) descrevem diversas técnicas para o problema de análise de agrupamento, criando uma classificação para estas. Por exemplo, duas classificações: técnicas hierárquicas, que possuem forte apelo gráfico (dendogramas) e técnicas aglomerativas, que são as mais conhecidas, como o *k-means* (KAUFMAN e ROUSSEEUW, 1990). Também indicam como estas técnicas podem ser estendidas para realizar detecção de *outliers* (valores discrepantes).

4.3. CLASSIFICAÇÃO E PREVISÃO

Os problemas de Predição estão relacionados à indução de modelos, pois suas técnicas buscam modelar uma variável chamada de dependente ou de resposta, através da relação desta com um conjunto de outras variáveis chamadas de explicativas, ou independentes. A ação modelar refere-se à busca da percepção do comportamento da variável, o que pode permitir realizar previsão (ou classificação) dos valores da variável dependente.

A indução de modelos é um problema antigo. Várias técnicas são desenvolvidas na busca de soluções mais eficientes. Por exemplo, as Redes Neurais existem desde 1943 (MCCULLOCH e PITTS, 1943), e a teoria estatística de decisão surge nos anos 1940 (PEARL, 1988), dentre outras. Destacam-se também as técnicas baseadas em árvores, que têm sido largamente desenvolvidas nos últimos vinte anos, desde os trabalhos de Breiman (1984) e Quinlan (1986), e se tornam populares. Diferentemente dos Modelos de Regressão Linear, que possuem fundamentalmente duas técnicas de estimação de seus parâmetros, diversas técnicas diferentes foram desenvolvidas usando estruturas de árvores de decisão.

Algumas técnicas de Classificação são descritas e comparadas em Coelho e Ebecken (2002), e em Coelho, Lachtermacher e Ebecken (2003) fazem um estudo semelhante considerando técnicas de previsão específicas, baseadas em árvores.

A distinção entre Classificação e Previsão, que são problemas resolvidos através da Indução de Modelos, é feita pela natureza da variável dependente. Dizemos que o modelo induzido é de Classificação quando são modeladas variáveis categóricas (assumem valores em um conjunto finito, sem ordenação natural, e suficientemente pequeno), e que é de Previsão quando são modeladas variáveis numéricas (assumem valores em um conjunto numérico, que por natureza é ordenado e é potencialmente infinito).

Alguns autores preferem o termo regressão ao invés de previsão. Outros discordam deste uso para o termo regressão. Esta diferenciação parede importante, pois modelos de regressão são uma conhecida técnica específica de modelagem de dados (é uma técnica de previsão) baseado na Teoria de Regressão Estatística. O uso deste termo pode então ser confuso e deve ser evitado.

5. FERRAMENTA DE GESTÃO

O principal uso das técnicas de Data Mining é, certamente, como Ferramentas de Gestão. O processo de descoberta de conhecimento em banco de dados pode trazer para o tomador de decisão informações importantes e surpreendentes. Estas informações permitem que sejam feitos posicionamentos estratégicos que não seriam considerados ou avaliados.

As aplicações de Data Mining como ferramenta de gestão são muitas e em diversas áreas. Para o presente texto, apenas as mais importantes aplicações diretamente relacionadas a negócios foram consideradas. Isto inclui aplicações para o Varejo e CRM – *Customer Relationship Management*, cuja diferença principal é a identificação do cliente (no caso do CRM). Consideramos também aplicações para Telecomunicação, Finanças (crédito e acompanhamento de mercados) e uma área que tem sido chamada de *Web Mining*. Algumas aplicações específicas do cenário nacional são consideradas com o intuito de mostrar que no mercado nacional também se usam estas técnicas. Pode-se considerar mais genericamente Carvalho (2001) ou Braga (2005).

As aplicações para varejo mais óbvias são baseadas em técnicas para Regras de Associação para Análise de Cesta de Mercado, como pode ser visto em Galindo, Coelho e Lachtermacher (2002), que usam dados de uma loja de conveniências da rede AmPm (Ipiranga). Goldschmidt e Passos (2005) falam sobre a aplicação da técnica numa cadeia de *Fast Food*. Por outro lado, Giudici (2003 p. 209-227) apresenta uma técnica baseada em regressão log-linear para investigar a associação entre as variáveis consideradas. Amaral (2001) traz algumas aplicações potenciais de Data Mining para marketing.

Berson, Smith & Thearling (1999) tratam especificamente de aplicações de algumas ferramentas de classificação e previsão, como redes neurais, árvores de decisão, KNN e Regressões Lineares para sistemas de CRM.

Giudici (2003, p. 273-291) apresenta um estudo de caso para uma empresa italiana de venda a distância, que quer diferenciar clientes com potencial para se tornar compradores leais dos clientes ocasionais. Este estudo pode identificar, num estágio inicial, quais clientes serão realmente lucrativos e quais devem ser tratados com um esforço concentrado de marketing para tornarem-se clientes leais. É um problema de Classificação, e é tratado usando técnicas de Regressão Logística, Redes Neurais (de base radial), Árvores de Classificação e KNN.

Goldschmidt e Passos (2005) mostram especificamente a questão do CRM para uma empresa de telefonia. O projeto tratava da classificação dos clientes segundo potencial de compra de serviços, para ações de marketing mais dirigidas e efetivas.

Westphal e Blaxton (1998, p. 465-475) apontam o uso de técnicas visuais (descritivas) de Data Mining para auxílio do problema de monitoramento de dados provenientes de sistemas de telecomunicações. Uma importante motivação para isto é a detecção de fraudes, que podem significar altos custos para a operadora. São poucas as aplicações de técnicas visuais tão bem sucedidas quanto esta. Outra aplicação mostrada pelos autores (p. 551-585), também sobre fraudes, mas na área financeira, usa ferramentas gráficas para detectar a lavagem de dinheiro.

Giudici (2003, p. 293-321) apresenta um estudo de caso para o problema conhecido como *credit scoring*, que se dá em termos de avaliar o crédito confiável a ser concedido para clientes (indivíduos) que solicitam crédito quando estão comprando bens ou serviços. São apresentadas as três soluções mais utilizadas para o problema: Regressão Logística, Arvore de Classificação e Redes Neurais, retropropagação (HAYKIN, 2000). Uma solução semelhante foi aplicada por Goldschmidt e Passos (2003) para o problema da caracterização de clientes que pagam em dia, que pagam em atraso ou clientes que não pagam os créditos, a partir de dados obtidos de uma financeira.

Uma série de atividades de Data Mining conduzida em um grande banco americano está descrita em Westphal e Blaxton (1998, p. 477-500). A previsão de crédito foi feita com os dados que incluíam localização geográfica. Os autores mostram ainda como ferramentas de

visualização puderam ser utilizadas para compreender melhor o modelo obtido.

Outra aplicação também descrita em Westphal e Blaxton (1998, p. 531-549) trata dos problemas vivenciados na gestão de carteiras de aplicações financeiras. O estudo foi conduzido considerando mercados financeiros asiáticos, e um dos resultados apontados pelos autores foi a previsão de um colapso em um banco japonês.

Um dos problemas que estão sendo tratados pelo grupo que se chama de *Web Miners* (analistas que usam técnicas de Data Mining no ambiente da *WWW*) é o comportamento do usuário da Internet. Giudici (2003 p. 230-253) mostra um estudo de caso do chamado *clickstream* – seqüência de seleções feitas com o mouse (cliques) por usuários em páginas de um determinado site. O objetivo da análise é descobrir, possivelmente *on-line*, quais páginas o usuário visitará, o que pode ser útil para sites dinâmicos. Como as informações sobre as seqüências de cliques são armazenadas nos servidores do site, os dados já estão disponíveis para análise. O autor usa técnicas de Regras de Associação e Cadeias de Markov, entre outras.

Uma outra análise sobre o comportamento do usuário de internet foi mostrada por Giudici (2003, p. 255-272). O objetivo era analisar dados de acesso à internet para identificar grupos de usuários com comportamento semelhante. Os dados utilizados foram provenientes do site da Microsoft, com mais de 32 mil visitantes anônimos. O autor usou duas técnicas de agrupamento: k-means e Redes Neurais, mapas de Kohonem (HAYKIN, 2000).

5.1. FERRAMENTAS COMPUTACIONAIS

As técnicas que foram desenvolvidas (e que continuam sendo) dão origem a um grande conjunto de ferramentas computacionais que são de grande utilidade para os tomadores de decisão em geral. Atualmente são muitas ferramentas disponíveis no âmbito comercial – sistemas que estão a venda e que têm empresas com respaldo garantindo serviços complementares de suporte, treinamento, etc. Existem também muitos sistemas no âmbito acadêmico – softwares gratuitos que foram desenvolvidos para testes e que podem ser utilizados a partir de pequenos investimentos financeiros ou até mesmo gratuitamente. Estes sistemas costumam ser eficientes do ponto de vista da qualidade da modelagem (precisão, custo computacional, interpretabilidade do modelo, etc.), mas costumam apresentar, via de regra, algum tipo de restrição nos aspectos escalabilidade ao tamanho da base de dados, interface ao usuário, documentação, suporte técnico, etc.

De uma maneira geral, estas ferramentas (tanto comerciais quanto acadêmicas) diferem-se nas técnicas e nos algoritmos de Data Mining implementados. Os algoritmos que já são bem conhecidos (como o Apriori – para Regras de Associação) estão disponíveis em diversas ferramentas. Neste caso, os softwares são diferenciados em pequenos detalhes da própria tecnologia, na interface com o usuário, ou na capacidade computacional.

Consideremos, por exemplo, a técnica de estimação de modelos de regressão linear baseada no método dos mínimos quadrados. Esta técnica é muito conhecida e está disponível em diversos softwares, desde simples planilhas eletrônicas como o Excel até sofisticadas suítes de Data Mining como o Enterprise Miner (SAS). Naturalmente, existem algumas diferenças em alguns detalhes da implementação da metodologia. Por exemplo, a ferramenta que se apresenta no Excel é bastante limitada para o volume de dados, o que não ocorre com a implementação do SAS. Isso não representa necessariamente uma deficiência do Excel, posto que as implementações foram feitas com propósitos diferentes. A metodologia para esta técnica de indução de modelos está absolutamente definida na literatura e não há, neste aspecto, nenhuma diferença entre as duas implementações que citamos e entre nenhuma outra.

Uma outra diferenciação que se faz sobre as ferramentas computacionais disponíveis para Data Mining é a diversidade de técnicas e algoritmos que a ferramenta oferece. Quando uma única ferramenta disponibiliza diversas opções de modelagens, ela é chamada de **suíte**. Uma suíte é geralmente uma ferramenta de KDD, no sentido que está desenhada para realizar várias etapas como pré-processamento dos dados ou visualização dos modelos obtidos. É o caso da suíte Clementine da SPSS, ou do Weka, da Universidade de Waikato.

6. METODOLOGIA DE PESQUISA

Temos presenciado um grande crescimento no volume da pesquisa empírica no meio acadêmico nos últimos tempos, principalmente na área de Administração. Isto pode ser facilmente observado através de uma observação sobre as recentes publicações nos principais congressos, jornais e revistas acadêmico-científicos. Grande parte destas pesquisas empíricas tem arcabouço quantitativo, baseado principalmente em estatística.

O método científico de comprovação empírica, seja quantitativo ou qualitativo, segue o princípio de que as observações de fatos reais devem corresponder ao modelo teórico, academicamente estabelecido. Por exemplo, a observação do comportamento de vendas de um certo produto em um certo canal de vendas frente a variações no preço estabelecido para este produto deve obedecer à Lei Microeconômica da Demanda que estabelece, em linhas gerais, que quanto maior o preço, menor a quantidade vendida (demandada).

Os métodos quantitativos de comprovação empírica são, em geral, baseados em estatística. Estes métodos estão, em última análise, relacionados a testes de hipóteses: as observações servem para buscar evidências para rejeitar ou não uma determinada afirmação – que pode ser um modelo teórico, academicamente aceito como a Lei de Demanda.

> Diferentemente das situações nas quais se podem empregar análises estatísticas ou matemáticas com a forma padrão para testar hipóteses pré-definidas, Data Mining é mais útil em cenários de análise exploratória nos quais não há noções pré-determinadas sobre o que constituirá um resultado interessante (WESTPHAL e BLAXTON, 1998, p. 6, tradução nossa).

Data Mining oferece uma maneira diferente de realizar a comprovação empírica. O seu método não está, e nem deve, diretamente relacionado com nenhuma hipótese previamente formalizada. O procedimento de descoberta do conhecimento deve ser conduzido de modo tão genérico que quaisquer padrões contidos nos dados sendo analisados possam ser percebidos. Estes padrões podem vir a ser hipóteses que corroborem com modelos teóricos ou não. Neste caso, o procedimento de descoberta do conhecimento se coloca como instrumento para ampliação de teorias existentes, que possam justificar os padrões descobertos.

> É importante manter em mente que o foco do processo de Data Mining é descobrir tendências e padrões ocultos [...] Entretanto, assim que um padrão particular é identificado, ele pode ser descrito como uma informação conhecida [...] Neste ponto, o processo de descoberta daquele padrão particular está acabado. [...] Abordagens analíticas que buscam nos conjuntos de dados pelos padrões conhecidos não estão realizando Data Mining, apesar de poderem estar usando os resultados dos processos de Data Mining para formar a base dos resultados das consultas. Por esta razão, nós não consideramos técnicas que requerem implementação de regras, exemplos de treinamento predefinidos ou aprendizado supervisionado automatizados como abordagens de Data Mining. Isto naturalmente não significa que tais técnicas não são úteis em muitos exemplos; isto simplesmente significa que, em nossa opinião, estes processos não constituem Data Mining (WESTPHAL e BLAXTON, 1998, p. 12, tradução nossa).

O uso de Data Mining como ferramenta de pesquisa requer do pesquisador uma postura diferente frente ao problema. O processo de comprovação empírica no sentido Teoria-Comprovação precisa ser modificado, para o sentido Conhecimento-Comprovação. O processo de comprovação passa pela validação do modelo de conhecimento obtido, e esta validação não pode ser feita, via de regra, com base nos mesmos dados que geraram o conhecimento.

6.1. CONSIDERAÇÕES ECONOMÉTRICAS

Existe na literatura de econometria uma grande resistência ao uso de Data Mining como método válido de pesquisa, apesar de ser frequentemente utilizado. Por exemplo, Sullivan et al (2001) mostram através de análise de dados financeiros o problema do viés quando não há uma teoria por traz do conhecimento obtido.

Entretanto, alguns econometristas tratam o problema de uma maneira diferente. De acordo com Backhouse e Morgan (2000), o uso de Data Mining é uma prática comum, apesar de ser considerada publicamente como indesejável. Pagan e Veall (2000, tradução nossa) afirmam que “ninguém duvidará que existe muita Data Mining na econometria”. Pagan e Veall (2000, tradução nossa) afirmam que “não há duvida de que existe muita Data Mining na

econometria".

Mayer concorda com esta idéia geral: "Data Mining é quase nunca definido, apenas raramente defendido, muito criticado, mas largamente praticado" (2000, tradução nossa), e aponta uma pesquisa entre economistas sobre o uso de Data Mining em resultados empíricos, onde 56% deles aceitam este procedimento traz alguma confiança sobre os resultados empíricos e 2% acham que o procedimento destrói qualquer confiança.

O problema parece residir no fato de que a prática de pesquisa econométrica está fundamentada com bastante tradicionalismo. Como conseqüência, os econometristas têm uma percepção muito estreita de como a pesquisa empírica pode ser conduzida. Por exemplo, o conceito dado para Data Mining por Hoover e Perez (2000, tradução nossa) é um tanto limitado: "'Data Mining' refere-se a uma gama de classes de atividades que tem em comum a busca de diferentes maneiras para processar ou sumarizar estatística ou econometricamente com o propósito de fazer a apresentação final se encaixar em determinado critério de projeto".

7. CONCLUSÃO

O desenho de uma taxonomia para as técnicas e algoritmos de Data Mining é importante por dois motivos. Para o usuário, uma taxonomia efetiva facilitaria o entendimento do cenário, permitindo uma compreensão rápida dos conceitos necessários evitando esforços desnecessários, no sentido permite uma pesquisa objetiva. Para o desenvolvedor – pesquisador que cria novas técnicas e algoritmos, a estruturação também traz vantagens.

A taxonomia proposta não está completa, e não é definitiva. Os principais problemas que Data Mining manipula foram considerados, mas existem outros apontados pela literatura e que não foram considerados. Um mapeamento das técnicas e suas relações com as funcionalidades também seria útil.

O uso de Data Mining como ferramenta de gestão não é mais uma novidade, mas existem ainda poucos trabalhos sendo desenvolvidos no cenário nacional, se comparado com o cenário mundial. Os esforços de divulgação são feitos em geral por pessoas isoladas, na forma de artigos de pesquisa ou estudos de casos, no máximo livros com teor teórico superficial. Os grandes fornecedores de solução em Data Mining (IBM, SAS, SPSS, etc.) não têm se posicionado ativamente frente a este mercado potencial.

Por outro lado, o uso de Data Mining como metodologia de pesquisa em administração é muito tímido. Possivelmente, os ruídos gerados por alguns econometristas tem tido um efeito sobre os pesquisadores de administração. Neste sentido, é importante considerar também os comentários positivos, dentre os quais citamos alguns. É necessário que o ambiente de pesquisa esteja apto a utilizar metodologias modernas, para que se tenha produção científica de qualidade e relevância.

8. REFERÊNCIAS

[1] Agrawal, R., Imielinksy, T. e Swami, A. **Mining Association Rules between Sets of Itens in Large Databases**. In: Proc. 1993 ACM SIGMOD Conference. Washington DC, EUA, 1993.

[2] Amaral, Fernanda Cristina Naliato do. **Data Mining: técnicas e aplicações para o marketing direto**. São Paulo: Berkeley, 2001. 128 p.

[3] Backhouse, Roger E. e Morgan, Mary S. **Introduction: is data mining a methodological problem?** Journal of Economic Methodology 7:2, (171-181), 2000.

[4] Berry, M. J. A.; Linoff, G. **Data Mining Techniques**. Nova York: John Wiley & Sons, 2004.

[5] Berson, Alex; Smith, Stephen; Thearling, Kurt. **Building Data Mining Applications for CRM**. Nova York, EUA: McGraw-Hill, 1999.510 p.

[6] Bittencourt, Guilherme. **Inteligência Artificial: ferramentas e teorias**. 2ª ed. Florianópolis: Editora da UFSC, 2001. 362 p.

[7] Braga, Luis Paulo V. **Introdução à Mineração de Dados**. Rio de Janeiro: E-papers, 2005.

[8] Breiman, L. et al. **Classification and Regression Trees**. Boca Raton, Florida, EUA: Chapman & Hall/CRC, 1984. 358 p.

[9] Carvalho, Luis Alfredo Vidal de. **Data Mining: a mineração de dados no marketing, medicina, economia, engenharia e administração**. São Paulo: Érica, 2001. 234 p.
[10] Coelho, P. S. S. e Ebecken, N. F. F. **A comparison of some classification techniques**. In: Data Mining III. Southampton (UK): Wit Press, 2002. p. 473-582.
[11] Coelho, P. S. S., Lachtermacher, G. e Ebecken, N. F. F. **A comparison of some tree baseed prediction tools**. In: Data Mining IV Southampton (UK): Wit Press, 2003.
[12] Date, C. J. **An Introduction to Database Systems**. 7. ed. Addison Wesley, 1999. 960 p.
[13] Galindo, M. G. L., Coelho, P. S. S. e Lachtermacher, G. **Usando Taxonomias para Criar Regras de Associação: aplicação em uma base de dados transacional**. In: Congresso Latino-Americano de Investigacion Operativa. Concepcion (CH). XI CLAIO, 2002.
[14] Goldschmidt, R. e Passos, E. **Data Mining: um guia prático**. Rio de Janeiro: Campus, 2005.
[15] Giudici, Paolo. **Applied Data Mining: statistical methods for business and industry**. West Sussex: John Wiley & Sons, 2003.
[16] Habermas, Jurgen. **Racionalidade e Comunicação**. Editora Edições 70, 2002
[17] Han, Jiawei; Kamber, Micheline. **Data Mining concepts and Techniques**. San Francisco: Morgan Kaufmann, 2001. 550 p.
[18] Haykin, S. S. **Redes Neurais: princípios e prática**. 2 ed. São Paulo: Bookman, 2000. 900 p.
[19] Hoover, Kevin D. e Perez, Stephen. **Three attitudes towards data mining**. Journal of Economic Methodology 7:2, (195-210), 2000.
[20] Kaufman, Leonard e Rousseeuw, Peter J. **Finding Groups in Data: an introduction to cluster analysis**. EUA: John Wiley & Sons, 1990.
[21] Kimball, R. e Ross, M. **The Data Warehouse Toolkit: guia completo para modelagem dimensional**. Trad. Tavares, Ana Beatriz e Lacerda, Daniela. Rio de Janeiro: Campus, 2002.
[22] Mayer, T. **Data mining: a reconsideration**. J. Economic Methodology 7:2, (183-194), 2000.
[23] McCulloch, W. S. e Pitts, W. **A logical calculus of the ideias immanent in nervous activity**. Bulletin of Mathematical Biophysics, vol. 5, pp. 115-133, 1943.
[24] Miller, Jerry P. e Business Intelligence Braintrust. **O milênio da inteligência competitiva**. Rubenich, Raul (trad.) Porto Alegre: Bookman, 2002. 293 p.
[25] Pagan, A. e Veall, M. **Data mining and the econometrics industry: comments on the papers of Mayer and Hoover and Perez**. J. of Economic Methodology 7:2, (211-16), 2000.
[26] Pearl, Judea. **Probabilistic Reasoning in Intelligent Systems: networks of plausible inference**. São Francisco, Califórnia, EUA: Morgan Kaufmann Publishers, Inc., 1998.
[27] Pyle, D. **Data Preparation for Data Mining**. Science & Technology Books, 1999. 504pp.
[28] Quinlan, John Ross. **Induction of Decision Trees**. Machine Learning, 1:81-106, 1986.
[29] Serra, Laercio. **A Essência do Business Intelligence**. São Paulo: Berkeley, 2002.
[30] Sullivan, R., Timmermann, A. e White, H. **Dangers of Data Mining: the case of calendar effects in stock returns**. Journal of Econometrics, 105 (249-286), 2001.
[31] Thomsen, Erik. **OLAP: construindo sistemas de informações multidimensionais**. Vieira, Daniel (trad.) Rio de Janeiro: Campus, 2002.
[32] Weiss, Sholom M.; Indurkhya, Nitin. **Predictive Data Mining: a pratical guide**. San Francisco (EUA): Morgan Kaufmann Publishers, 1998. 228 p.
[33] Westphal, Cristopher e Blaxton, Teresa. **Data Mining Solutions: methods and tools for solving real-world problems**. John Wiley & Sons, 1998
[34] Witten, Ian H. & Frank, Eibe. **Data Mining: Practical machine learning tools with Java implementations**. 2a ed. San Francisco: Morgan Kaufmann (Elsevier), 2005.