UEM

MINERAÇÃO DE DADOS (*DATA MINING*)

Profa. Da. Maria Madalena Dias

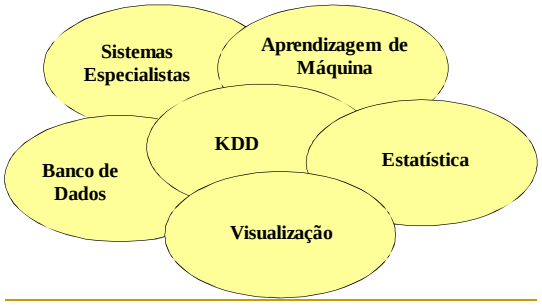
Tópicos

- O que é Mineração de Dados
- Mineração de Dados como um Campo Multidisciplinar
- Problemas de Mineração de Dados
- Mineração de Dados X Aprendizagem de Máquina
- Mineração de Dados X Estatística
- O Processo de KDD (*Knowledge Discovery in Data Base*)
- Tarefas de Mineração de Dados
- Técnicas de Mineração de Dados
- Áreas de Aplicação da Mineração de
- Aplicação de Técnicas de Mineração de Dados
- Visualização
- Mineração Visual de Dados
- Aplicação de Técnicas de Visualização
- Conclusões

O que é 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, p.5).
- Tem como principais objetivos:
  - Descobrir relacionamentos entre dados;
  - Fornecer subsídios para que possa ser feita uma previsão de tendências futuras baseando-se no passado.

Mineração de Dados como um Campo Multidisciplinar



(Cratochvil, 1999)

Problemas de mineração de dados

- Número de relacionamentos possíveis é muito grande
  - necessidade de estratégias de busca inteligentes como Aprendizagem de Máquina.
- Informações corruptas ou perdidas
  - necessidade de aplicação de técnicas estatísticas para estimar a confiabilidade dos relacionamentos descobertos.

(Holsheimer; Siebes, 2004)

Mineração de Dados X Aprendizagem de Máquina

MINERAÇÃO DE DADOS	APRENDIZAGEM DE MÁQUINA
Tipo especial de aprendizagem de máquina no qual o ambiente é observado através de um banco de dados.	Automação de um processo de aprendizagem .
O conjunto de treinamento é um banco de dados que contém outros tipos de dados além dos numéricos.	O conjunto de treinamento contém dados numéricos.
O sistema não pode manipular seu ambiente para gerar exemplos interessantes.	O sistema tem habilidade para interagir com seu ambiente.

(Holsheimer; Siebes, 2004)

Mineração de Dados X Estatística

MINERAÇÃO DE DADOS

- Dados numéricos ou não.
- Grandes volumes de dados .
- Os dados tendem a ser ruidosos e os valores para os atributos são frequentemente omitidos.
- Necessidade de dados serem processados em tempo real.

ESTATÍSTICA

- Apenas dados numéricos.
- Volume de dados limitado.
- Técnicas estatísticas são usadas para lidar com valores ruidosos e omitidos.
- Não requerem processamento em tempo real.

(Hand, 1999; Hand, 2004)

Mineração de Dados X Estatística

MINERAÇÃO DE DADOS

- O programa impulsiona a análise, pois o usuário tem recursos insuficientes para examinar manualmente bilhões de registros e centenas de milhares de padrões potenciais .
- Ajuda o usuário na geração de hipóteses.
- Em muitas situações, todos os dados possíveis estão disponíveis, e o objetivo não é fazer inferência, mas sim, descrever esses dados.

ESTATÍSTICA

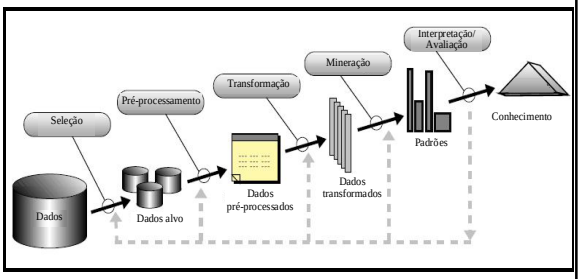
- O programa impulsiona a análise a seguir uma estratégia de estatística, pois o usuário tem conhecimento estatístico insuficiente para fazê-lo.
- Hipóteses são formuladas pelo usuário.
- Preocupa-se com a inferência de uma forma ou de outra, o objetivo é usar os dados disponíveis para fazer declarações sobre a população da qual os dados foram retiradas.

(Hand, 1999; Hand, 2004)

O Processo de KDD

- "*KDD*" ou Descoberta de Conhecimento em Banco de dados refere-se ao processo que tem por objetivo extrair conhecimento em grandes volumes de dados, com a aplicação de técnicas de Mineração de dados
- O processo de KDD é composto por três etapas principais: pré-processamento, mineração de dados e pós-processamento

O Processo de KDD



(Fayyad, 1996)

O Processo de KDD

Definição de Objetivos	→ Definição de objetivos de negócio a alcançar → Planejamento de avaliação de resultados
Preparação de Dados	→ Seleção e transformação dos dados → Registro no metadados → Registro no data warehouse, data mart ou data set
Definição de um Estudo	→ Articular um alvo para estudo → Escolher uma variável dependente → Especificar os campos de dados usados no estudo
Construção de um Modelo	→ Aplicação de uma técnica de mineração de dados
Análise do Modelo	→ Análise baseada no tipo de técnica de mineração de dados aplicada → Uso de técnicas de visualização
Predição	→ Escolha do melhor resultado possível

(Dias, 2001)

Processo de KDD

- PRÉ-PROCESSAMENTO
  - Uso de técnicas estatísticas para análise dos dados existentes com o objetivo de eliminar dados incompletos, ruidosos ou inconsistentes e definir a discretização.
  - Existência ou não de uma estrutura de Data Warehouse ou organização dos dados em uma única tabela – junção direta ou junção orientada.
  - Redução de dados horizontal: amostragem aleatória, eliminação direta de casos, segmentação do banco de dados e agregação de informações.
  - Redução de dados vertical (redução de dimensão): eliminação ou substituição dos atributos de um conjunto de dados.
    - Abordagem independente do modelo (*Filter*): não considera o algoritmo de MD a ser aplicado;
    - Abordagem dependente do modelo (*Wrapper*): experimentar o algoritmo de MD para cada conjunto de atributos e avaliar os resultados obtidos.

(Goldschmidt e Passos, 2005)

Processo de KDD

- PRÉ-PROCESSAMENTO
  - Redução de Valores:
    - Redução de valores nominais: identificação de hierarquia entre atributos e identificação de hierarquia entre valores;
    - Redução de valores contínuos (ou discretos): pressupõe aplicação somente a variáveis contínuas e discretas.
  - Limpeza: envolve uma verificação da consistência das informações, a correção de possíveis erros e o preenchimento ou a eliminação de valores desconhecidos e redundantes, além da eliminação de valores não pertencentes ao domínio.
  - Codificação:
    - Numérica – Categórica: divide valores de atributos contínuos em intervalos codificados;
    - Categórica – Numérica: representa valores de atributos categóricos por códigos numéricos.

(Goldschmidt e Passos, 2005)

Processo de KDD

- MINERAÇÃO DE DADOS
  - Técnicas baseadas em aprendizagem de máquina, técnicas estatísticas e inteligência artificial.
  - Identificação da tarefa de mineração de dados de acordo com o alvo de estudo e o tipo de problema a ser solucionado;
  - Definição da técnica e do algoritmo de mineração de dados a ser aplicada.

Processo de KDD

- PÓS-PROCESSAMENTO
  - Resultados: regras

Tarefas de Mineração de Dados

- Classificação
- Estimativa (ou Regressão)
- Associação
- Segmentação (ou *Clustering*)
- Sumarização

Técnicas de Mineração de Dados

- Descoberta de Regras de Associação
- Árvores de Decisão
- Raciocínio Baseado em Casos
- Algoritmos Genéticos
- Redes Neurais Artificiais

Descoberta de Regras de Associação

- Estabelece uma correlação estatística entre atributos de dados e conjuntos de dados
- Tem a forma geral X1 ^ ... ^ Xn => Y [C,S], onde X1,..., Xn são atributos que prevêem a ocorrência de Y com um grau de confiança C e com um suporte mínimo de S
- Tarefa:
  - Associação
- Exemplo:
  - determinar quais produtos costumam ser colocados juntos em um carrinho de supermercado.

Árvores de Decisão

- Hierarquização dos dados, baseada em estágios de decisão (nós) e na separação de classes e subconjuntos
- Tarefas:
  - Classificação
  - Estimativa
- Exemplos:
  - classificar pedidos de créditos como de baixo, médio e alto risco;
  - esclarecer pedidos de seguros fraudulentos;
  - identificar a melhor forma de tratamento de um paciente;
  - prever a demanda de um consumidor para um novo produto.

Raciocínio Baseado em Casos

- Baseado no método do vizinho mais próximo, combina e compara atributos para estabelecer hierarquia de semelhança
- Tarefas:
  - Classificação
  - Segmentação
- Exemplos:
  - agrupar clientes por região do país;
  - agrupar clientes com comportamento de compra similar;
  - agrupar seções de usuários Web para prever comportamento futuro de usuário.

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"
- Tarefas:
  - Classificação
  - Segmentação

Redes Neurais Artificiais

- Modelos inspirados na fisiologia do cérebro humano, onde o conhecimento é fruto do mapa das conexões neuronais e dos pesos dessas conexões
- Tarefas:
  - Classificação
  - Segmentação

Aplicação de Técnicas de Mineração de Dados

- Áreas de aplicação:
  - Marketing
  - Detecção de fraudes
  - Instituições governamentais
  - Controle de processos e de qualidade
  - Transporte
  - Banco
  - Ciência
  - Apólice de seguro e cuidado da saúde
  - Medicina
  - C & T (Ciência e Tecnologia)
  - Web

Aplicação de Técnicas de Mineração de Dados

- Weka
- C & T (Ciência e Tecnologia)
  - Cursos de pós-graduação brasileiros
- Detecção de Fraudes
  - Clientes de cartão de crédito
- Instituição Governamental
  - Contribuintes municipais

WEKA

- *Waikato Environment for Knowledge Analysis*
- É um pacote Java desenvolvido em uma universidade da Nova Zelandia
- Técnicas de mineração de dados:
  - Árvore de decisão (classificação) – J48
  - Regras de associação - Apriori
  - Raciocínio baseado em caso (*clustering*) - K-médias
- http://www.cs.waikato.ac.nz/ml/weka

Formato do arquivo de entrada do Weka

```
% Isso é um comentário
@relation BaseORIENTADORFORMANDO
@attribute AREA
    {OUTROS,CIÊNCIASEXATASEDATERRA,CIÊNCIASBIOLÓG
    ICAS,ENGENHARIAS,CIÊNCIASHUMANAS,CIÊNCIASSOCI
    AISAPLICADAS,CIÊNCIASAGRÁRIAS,CIÊNCIASDASAÚDE,
    LINGUÍSTICALETRASEARTES}
@attribute NUMORIENTANDOS {1,2,3,7,8}
@attribute MESESFORMACAO {12,18,24,30,31}

@data
CIÊNCIASHUMANAS,7,31
CIÊNCIASHUMANAS,3,31
CIÊNCIASHUMANAS,7,31
CIÊNCIASEXATASEDATERRA,2,31...
```

C&T
Análise dos Resultados - J48

- Relação entre quantidade de bolsas fornecidas ao Programa e a produtividade do seu corpo docente
  - 61% dos pesquisadores produzem em média 2 a 3 publicações por ano
  - 74% dos pesquisadores da área de Lingüística, Letras e Artes e 50% da área de Ciências Humanas produzem em média 4 a 7 publicações por ano

C&T
Análise dos Resultados – Apriori

- Relação entre quantidade de bolsas fornecidas ao Programa e a produtividade do seu corpo docente
  - 60% dos pesquisadores produzem 2 a 3 publicações em média por ano
  - 70% dos programas cujos alunos recebem em média 19 a 24 meses de bolsa, seus pesquisadores produzem 2 a 3 publicações em média por ano
  - 56% dos programas cujos alunos recebem em média 25 a 30 meses de bolsa, seus pesquisadores produzem 2 a 3 publicações em média por ano

C&T
Análise dos Resultados – J48

- Relação entre a carga de orientação e o tempo de formação dos orientandos
  - 73% dos alunos de mestrado demoram mais de 30 meses para se formar, independente do número de orientandos que seu orientador possui

C&T
Análise dos Resultados – Apriori

- Relação entre a carga de orientação e o tempo de formação dos orientandos
  - 77% dos alunos, cujo orientador possui de 4 a 7 orientandos, demoram mais de 30 meses para se formar
  - 73% dos alunos, cujo orientador possui 3 orientandos, demoram mais de 30 meses para se formar
  - 71% dos alunos, cujo orientador possui 1 orientando, demoram mais de 30 meses para se formar
  - 69% dos alunos, cujo orientador possui 2 orientandos, demoram mais de 30 meses para se formar

C&T
Análise dos Resultados – J48

- Relação entre o tempo de titulação com a disponibilidade de bolsas no Programa ou com a participação discente em projetos
  - 70% dos alunos de mestrado demoram mais de 30 meses para se formar, independente de ter ou não bolsa ou de sua dissertação estar ou não vinculada com algum projeto
  - 68,5% das dissertações possuem vínculo com projeto

C&T
Análise dos Resultados - Apriori

- Relação entre o tempo de titulação com a disponibilidade de bolsas no Programa ou com a participação discente em projetos
  - 75% dos alunos que possuem bolsa, suas dissertações possuem vínculo com projeto
  - 70% dos alunos que possuem bolsa demoram mais de 30 meses para se formar
  - 69% dos alunos, cuja dissertação possui vínculo com projeto, demoram mais de 30 meses para se formar

C&T
Análise dos Resultados - K-médias

- Meses de formação
  - Estabilização dos clusters em 3 grupos:
    - 80% - mais de 30 meses
    - 19% - 25 a 30 meses
    - 1% - 1 a 12 meses
- Número orientandos
  - Formação de 4 clusters
    - Centros em 2 (25%), 7 (32%), 1 (22%) e 3 (21%) orientandos

C&T
Análise dos Resultados - K-médias

- Formação de 4 clusters com meses de formação e número de orientandos
  - 58% dos alunos, cujo orientador possui 2 orientandos, demoram mais de 30 meses para se formar
  - 12% dos alunos, cujo orientador possui 7 orientandos, demoram de 25 a 30 meses para se formar
  - 25% dos alunos, cujo orientador possui 7 orientandos, demoram mais de 30 meses para se formar
  - 6% dos alunos, cujo orientador possui 3 orientandos, demoram de 25 a 30 meses para se formar

C&T
Conclusões

- A produtividade do pesquisador está mais relacionada à área que ele pertence do que à quantidade de bolsa fornecida ao Programa de Pós-Graduação que ele pertence
- O tempo de formação do aluno de mestrado não está diretamente relacionado com: o número de orientandos de seu orientador, o fato de possuir ou não bolsa e a sua dissertação estar ou não vinculada a projeto

C&T
Conclusões

- As tarefas de classificação e associação podem ser consideradas adequadas na descoberta de conhecimento da área de C&T, os resultados obtidos são bastante semelhantes
- O uso da tarefa de segmentação (*clustering*) é mais adequado na fase de preparação de dados, pelo fato da mesma não fazer correlação entre dados

Detecção de Fraudes

- **Status da Fatura → Característica do Cliente**
  - Fatura vencida em até 30 dias → Bom Pagador
  - Fatura vencida entre 31 e 180 dias → Devedor
  - Fatura vencida acima de 181 dias → Mau Devedor
  - Fatura paga em até 30 dias → Bom Pagador
  - Fatura paga entre 31 e 180 dias → Bom Devedor
  - Fatura paga acima de 181 dias → Mau Pagador

Detecção de Fraudes

- Atributos:
  - Situação do Cliente (Atributo Meta)
  - Faixa Etária
  - Sexo
  - Estado Civil
  - Número de Dependentes
  - Salário
  - Cidade
  - Bairro
  - Cargo

Detecção de Fraudes
Análise dos Resultados – J48

- 77,68% dos clientes são Bons Pagadores
- O atributo cargo oferece maior influência sobre os clientes com status Bom Pagador.
  - Mais de 80% dos clientes com cargo de Cozinheiro, Comerciário, Cabeleireiro, Estagiário, Aposentado, Pensionista, Operador, Serviços Gerais, Auxiliar, Escritório, Costureira, Motorista e Assistente de Obras são Bons Pagadores.

Detecção de Fraudes
Análise dos Resultados – J48

- O atributo bairro também exerce grande influência:
  - 80% dos clientes que moram nas zonas 1 e 21 são Bons Pagadores
  - 80% dos clientes que moram na zona 54 e possuem o cargo Outros são Mau Devedores
  - 75% dos clientes que moram na zona 29 são Mau Devedores
  - 71% dos clientes que moram em Paiçandu e possuem menos de 20 anos são Mau Devedores
  - 56% dos clientes que moram na zona 43 e possuem idade menor ou igual a 25 anos são Mau Devedores

Detecção de Fraudes
Análise dos Resultados – J48

- 56% dos clientes que moram nas zonas 37 e 38 e possuem cargo Outros e idade 21 a 25 anos são Mau Devedores
- 50% dos clientes que moram em Sarandi e possuem de 36 a 40 anos são Mau Devedores
- 44% dos clientes que moram em Paiçandu e possuem o cargo Outros são Mau Devedores
- 40% dos clientes que moram em Sarandi e possuem menos de 20 anos e de 31 a 35 anos são Mau Devedores

Detecção de Fraudes
Análise dos Resultados – J48

- O atributo idade também exerce influência sobre o status Bom Pagador:
  - 77% de clientes com idade entre 26 e 30 anos são Bons Pagadores
  - 73% dos clientes com idade menor ou igual a 25 anos são Bons Pagadores

Detecção de Fraudes
Análise dos Resultados – J48

- Os clientes de outros municípios são os que apresentam status de Mau Pagador
  - 44% dos clientes que moram em outros municípios, possuem de 21 a 25 anos e cargo igual a Outros são Mau Pagadores
  - 43% dos clientes que moram em outros municípios e possuem de 26 a 30 anos são Mau Pagadores

Detecção de Fraudes
Análise dos Resultados – K-médias

- Formação de 2 clusters:
  - 49% - Bom Pagador, com idade entre 21 a 25 anos, sexo masculino, residente em Sarandi, casado, sem filhos e cargo serviços gerais
  - 51% - Bom Pagador, idade abaixo de 20, sexo feminino, residente em Maringá, solteira e cargo serviços gerais

Detecção de Fraudes
Conclusões

- Os atributos cargo, bairro e idade são os que mais influenciam no perfil do cliente bom pagador
- O atributo bairro e cidade são os que mais influenciam no perfil do cliente mau devedor
- O atributo cidade é o que mais influencia no perfil do cliente mau pagador
- A tarefa de classificação é a mais adequada para a área de detecção de fraude, considerando que o principal objetivo é descobrir o perfil do bom cliente e do inadimplente

Instituições governamentais
Análise dos Resultados – J48

- 70% dos contribuintes são Bons Pagadores
- mais de 95% dos contribuintes que possuem imóveis nas zonas 1 e 16 são Bons Pagadores
- 90% dos contribuintes que possuem imóveis nas zonas 14 e 28 são Bons Pagadores
- 76 a 81% dos contribuintes que possuem imóveis nas zonas 2, 9, 11 e 21 são Bons Pagadores
- 70% dos contribuintes que possuem imóveis nas zonas 3, 4 e 7 são Bons Pagadores
- os Mau Pagadores se concentram nas zonas 32, 30, 31, 23, 24 e 29, com respectivas inadimplências 84%, 80%, 79%, 70%, 69% e 56%.

Instituições governamentais
Análise dos Resultados – K-médias

- Formação de 2 clusters:
  - 29% - Mau Pagador com imóvel na zona 07
  - 71% - Bom Pagador com imóvel na zona 01
- Formação de 5 clusters:
  - 22% - Mau Pagador com imóvel na zona 07
  - 37% - Bom Pagador com imóvel na zona 07
  - 5% - Bom Pagador com imóvel na zona 16
  - 34% - Bom Pagador com imóvel na zona 01
  - 3% - Bom Devedor com imóvel na zona 01

Instituições governamentais
Conclusões

- A tarefa de classificação apresenta melhores resultados do que a tarefa de segmentação
- Os resultados do algoritmo K-médias não possibilita uma análise de percentual de inadimplência por bairro

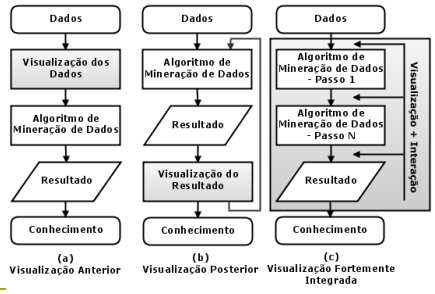
Universidade (vestibular)
Análise dos Resultados – J48

- Base de dados contendo dados dos candidatos aos onze cursos **mais** concorridos
  - Nota de redação exerce grande influência na classificação do candidato
  - Notas da língua portuguesa e redação, em conjunto com as notas de matemática e química, influenciam na aprovação do candidato
  - Notas do ENEM acima de 42 pontos indicam a possível aprovação do candidato
  - Nota de química, juntamente com as notas de Língua Estrangeira e Língua Portuguesa, influenciou no resultado dos candidatos ao curso de Direito

Universidade (vestibular)
Análise dos Resultados – J48

- Base de dados contendo dados dos candidatos aos onze cursos **menos** concorridos
  - Notas de Geografia, Redação, Língua Estrangeira e História influenciam, em geral, no resultado do vestibular
  - Notas de Matemática, Química e Física não exercem tanta influência no resultado
  - Grande número de candidatos aprovados que concluíram o ensino médio há mais de 5 anos
  - O fato do candidato ter ou não feito cursos pré-vestibulares não exerce grande influência no resultado do vestibular

Integração de Visualização com o Processo de KDD

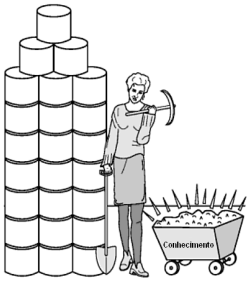


(Anskert, 2001)

Universidade (vestibular)
Análise dos Resultados – J48

- Base de dados contendo dados de todos os candidatos ao vestibular
  - A maioria dos candidatos não trabalha, reside com os pais e se encontra na faixa etária de 17 a 20 anos
  - Notas de Geografia, Língua Estrangeira e História influenciam na aprovação do candidato nos cursos da área de Exatas
  - As notas de Matemática, Física e Química contribuem para a aprovação do candidato nos cursos da área de Humanística
  - A nota de Língua Estrangeira, em conjunto com Geografia e História, influencia diretamente na aprovação do candidato

Visualização

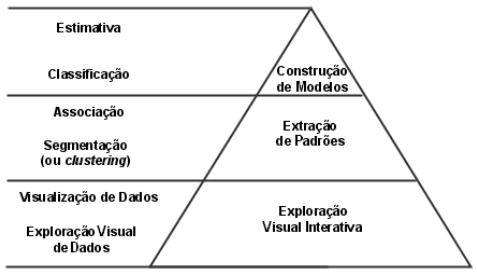


- Processo de KDD
- Algoritmos de mineração de dados
- Técnicas de visualização

(Han e Kamber, 2001)

Visualização

- Níveis de sistemas de mineração



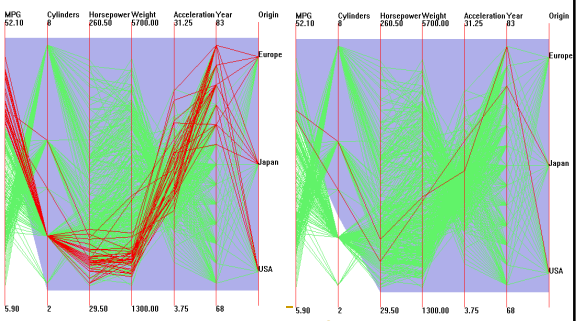
(Mendonça; Sunderhaft, 1999)

Mineração Visual de Dados

- Refere-se à aplicação de técnicas de visualização para avaliar, monitorar e guiar o processo de mineração de dados.
  - Essa avaliação consiste de exemplos de treinamento, de teste e modelos de aprendizado para verificar os resultados dos algoritmos de mineração (Ganesh et al.,1996),
  - O monitoramento inclui atividades como rastrear o progresso dos algoritmos de mineração de dados, avaliando a relevância dos padrões no contexto das atualizações sofridas pelo banco de dados.

Aplicação de Técnicas de Visualização

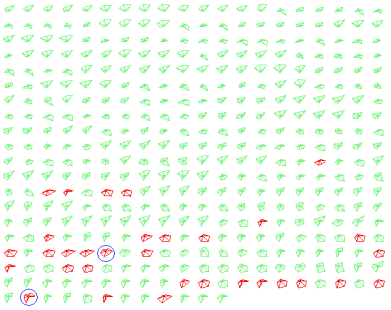
Base de Dados Automóveis - Coordenadas Paralelas



Ferramenta Xmdv

Aplicação de Técnicas de Visualização

Base de Dados Automóveis – Star Glyphs



Ferramenta Xmdv

Conclusões

- O usuário de um sistema KDD 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;
- A etapa de pré-processamento é geralmente bastante trabalhosa e toma muito tempo na descoberta de conhecimento em banco de dados;
- O ideal é sempre ter um DW construído para facilitar a aplicação de técnicas de mineração de dados;

Conclusões

- Mineração de dados não será descoberta de conhecimento sem estatística.
- Estatística não será capaz de ter sucesso em conjuntos de dados maciços e complexos sem abordagens de mineração de dados.

(Kuonen, 2004)

Referências Bibliográficas

- BERRY, M.J.A.; LINOFF, G. Data mining techniques. John Wiley & Sons, Inc. 1997.
- CRATOCHVIL, A. Data mining techniques in supporting decision making. Master thesis, Universiteit Leiden, 1999.
- HAND, D.J. Why data mining is more than statistics writ large. *Bulletin of the International Statistical Institute*, 52nd Session, Vol. 1, 1999, p. 433-436.
- HAND, D.J. Data mining: statistics and More? *The American Statistician*, May 1998 Vol. 52, No. 2, p. 112-118.

Referências Bibliográficas

- HOLSHEIMER, M.; SIEBES, A. The search for knowledge in databases. Report C-R9406, ISSN 0169-118X, Amsterdam, The Netherlands, 2004.
- KUONEN, D. Data mining and statistics: what is the connection? *TDAN.com October*, 2004 .

Conclusões

- Existem ferramentas específicas para construção de DW que oferecem alguns recursos para busca de informações sem a aplicação de técnicas de mineração de dados;