ACADÊMICO: GEANDRO LUÍS COMPOLT

ORIENTADOR: OSCAR DALFOVO

Sistema de Informação Executiva Baseado em Data Mining Utilizando a Técnica de Árvores de Decisão

Roteiro

- Introdução;
- Sistemas de Informação;
- Data Mining;
- Desenvolvimento do protótipo e do SIE;
- Conclusões;
- Sugestões.

Introdução

- **Motivação**
  - – possibilitar as empresas aproveitar de forma mais eficaz as informações que estão armazenadas em seus arquivos;
  - – possibilitar aos executivos novas formas de visualização e compreensão das informações inerentes ao seu negócio.

- **Objetivo**

  Auxiliar o processo de tomada de decisões de uma empresa, através de um Sistema de Informação Executiva utilizando técnicas de Data Mining, mais especificamente para efetuar classificações e segmentações.

Sistemas de Informação

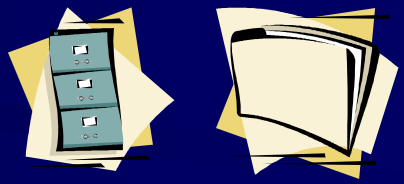
Conceito

Sistema especializado que pode ser definido por um conjunto de elementos ou componentes inter-relacionados que coletam (entrada), manipulam e armazenam (processo) disseminam os dados e informações (saída) e fornecem um mecanismo que permitem realizar ajustes ou modificações no nas atividades de entrada ou processamento (feed-back).

Elementos de um Sistema de Informação

Tecnologia da Informação

Pessoas



Informação

Técnicas

Objetivos

Data Mining

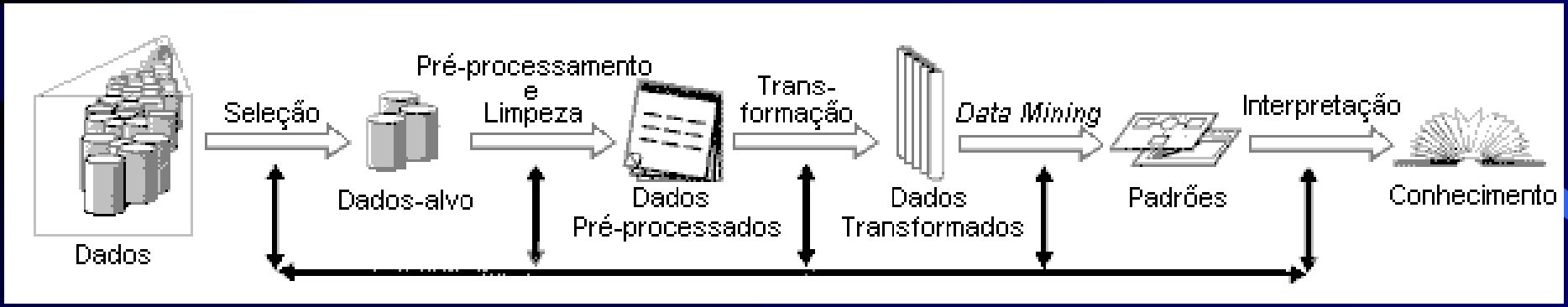
- **Conceito**

  É a exploração e análise, por meios automáticos ou semi-automáticos, de uma grande quantidade de dados para descobrir padrões e regras significativas [BER97].

Data Mining (continuação)

- **KDD (Prospecção de conhecimento em bases de dados)**

  Processo que envolve a automação da identificação e do reconhecimento de padrões em um banco de dados.



Passos do processo de KDD

Data Mining (continuação)

- **Funções**
  - classificação;
  - estimativa;
  - agrupamento por afinidade;
  - previsão;
  - segmentação.

Árvores de Decisão

O objetivo desta técnica é reconhecer de forma automática a representação de formas simples de lógica condicional buscando a representação de uma série de questões que estão escondidas sobre a base da dados formando assim uma estrutura em árvore.

Em uma árvore de decisão existem dois tipo de atributos; o decisivo, que contém o resultado ou alvo ao qual se quer atingir e os não decisivos que contém os valores que conduzem a uma tomada de decisão [QUI93].

Entendimento da técnica

NOME	FAT	VOL	D.VCER	D.CIDA	METAS	LIMCRE	CONJ	SPC	AD
Alberto Reis	100000-500000	MEDIO	NÃO	NÃO	NÃO	ALTO	SIM	SIM	SIM
Caio de Abreu	0-10000	BAIXO	NÃO	NÃO	NÃO	MEDIO	NAO	NAO	SIM
Castelo Branco	100000-500000	BAIXO	NÃO	NÃO	NÃO	ALTO	SIM	NÃO	SIM
Claudio Tafarel	ACIMA 500000	BAIXO	SIM	NÃO	SIM	BAIXO	SIM	SIM	NAO
Jardel de Souza	0-100000	MEDIO	NÃO	SIM	NÃO	ALTO	SIM	SIM	NAO
Jose da Silva	0-100000	ALTO	NÃO	NÃO	NÃO	ALTO	NAO	SIM	NAO
Pedro de Assis	ACIMA 500000	MEDIO	SIM	SIM	SIM	BAIXO	SIM	SIM	NAO

$$ENTROPIA(S) := \sum -p(I) * \log_2 p(I)$$

$$ENTROPIA(S) := (-3/7 * \log_2 3/7) + (-4/7 * \log_2 4/7)$$

$$ENTROPIA(S) := (-3/7 * -1.222) + (-4/7 * -0.807)$$

$$ENTROPIA(S) := 0.985$$

Entendimento da técnica (continuação)

NOME	FAT	VOL	D.VCER	D.CIDA	METAS	LIMCRE	CONJ	SPC	AD
Alberto Reis	100000-500000	MEDIO	NÃO	NÃO	NÃO	ALTO	SIM	SIM	SIM
Castelo Branco	100000-500000	BAIXO	NÃO	NÃO	NÃO	ALTO	SIM	NÃO	SIM
Caio de Abreu	0-10000	BAIXO	NÃO	NÃO	NÃO	MEDIO	NAO	NAO	SIM
Jardel de Souza	0-10000	MEDIO	NÃO	SIM	NÃO	ALTO	SIM	SIM	NAO
Jose da Silva	0-10000	ALTO	NÃO	NÃO	NÃO	ALTO	NAO	SIM	NAO
Claudio Tafarel	ACIMA 500000	BAIXO	SIM	NÃO	SIM	BAIXO	SIM	SIM	NAO
Pedro de Assis	ACIMA 500000	MEDIO	SIM	SIM	SIM	BAIXO	SIM	SIM	NAO

$$Gain(S,A) = Entropia(S) - \sum ((|Sv|) / |S|) * Entropia(Sv))$$

$$Gain(FAT,A) = 0.985 - ((2/7) * 0 + (3/7) * 0.918 + (2/7) * 0)$$

$$Gain(FAT,A) = 0.591$$

Entendimento da técnica (continuação)

NOME	FAT	VOL	D.VCER	D.CIDA	METAS	LIMCRE	CONJ	SPC	AD
Alberto Reis	100000-500000	MEDIO	NÃO	NÃO	NÃO	ALTO	SIM	SIM	SIM
Castelo Branco	100000-500000	BAIXO	NÃO	NÃO	NÃO	ALTO	SIM	NÃO	SIM
Caio de Abreu	0-10000	BAIXO	NÃO	NÃO	NÃO	MEDIO	NAO	NAO	SIM
Jardel de Souza	0-10000	MEDIO	NÃO	SIM	NÃO	ALTO	SIM	SIM	NAO
Jose da Silva	0-10000	ALTO	NÃO	NÃO	NÃO	ALTO	NAO	SIM	NAO
Claudio Tafarel	ACIMA 500000	BAIXO	SIM	NÃO	SIM	BAIXO	SIM	SIM	NAO
Pedro de Assis	ACIMA 500000	MEDIO	SIM	SIM	SIM	BAIXO	SIM	SIM	NAO

Gain (FAT,A) = 0.591

Gain (VOL,A) = 0.198

Gain (D.VCER,A) = 0.291

Gain (D.CIDA,A) = 0.291

Gain (METAS,A) = 0.291

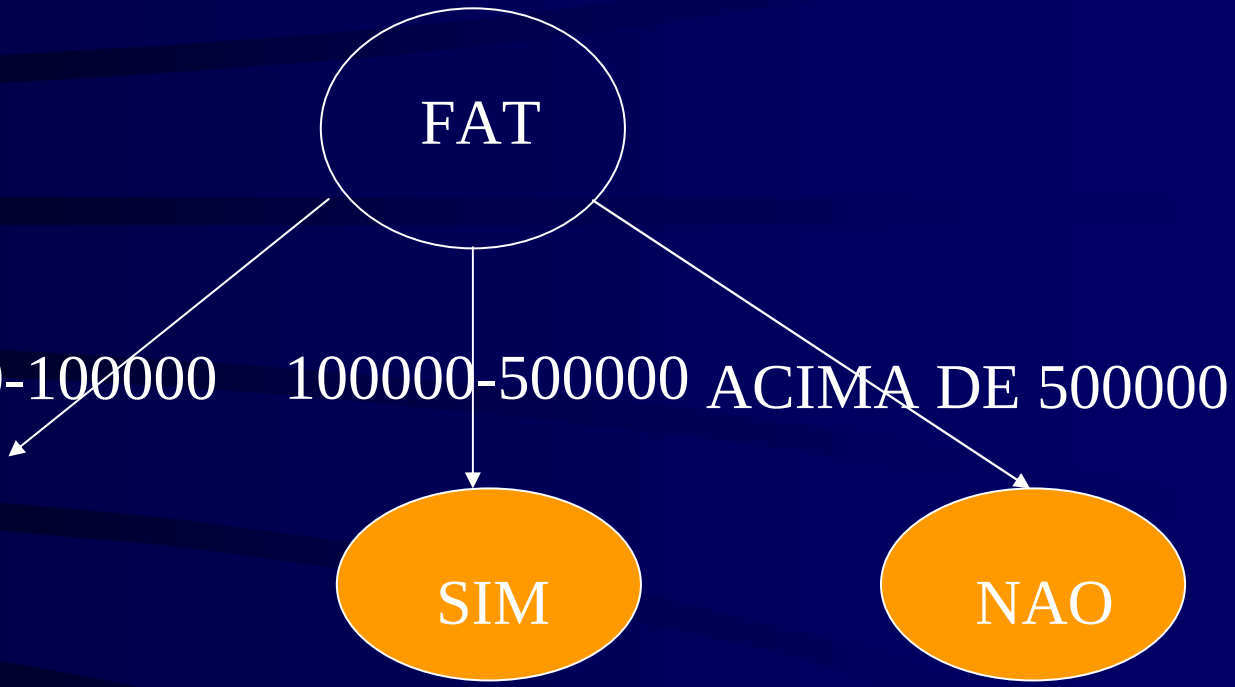
Gain (LIMCRE,A) = 0.413

Gain (CONJ,A) = 0.005

Gain (SPC,A) = 0.469

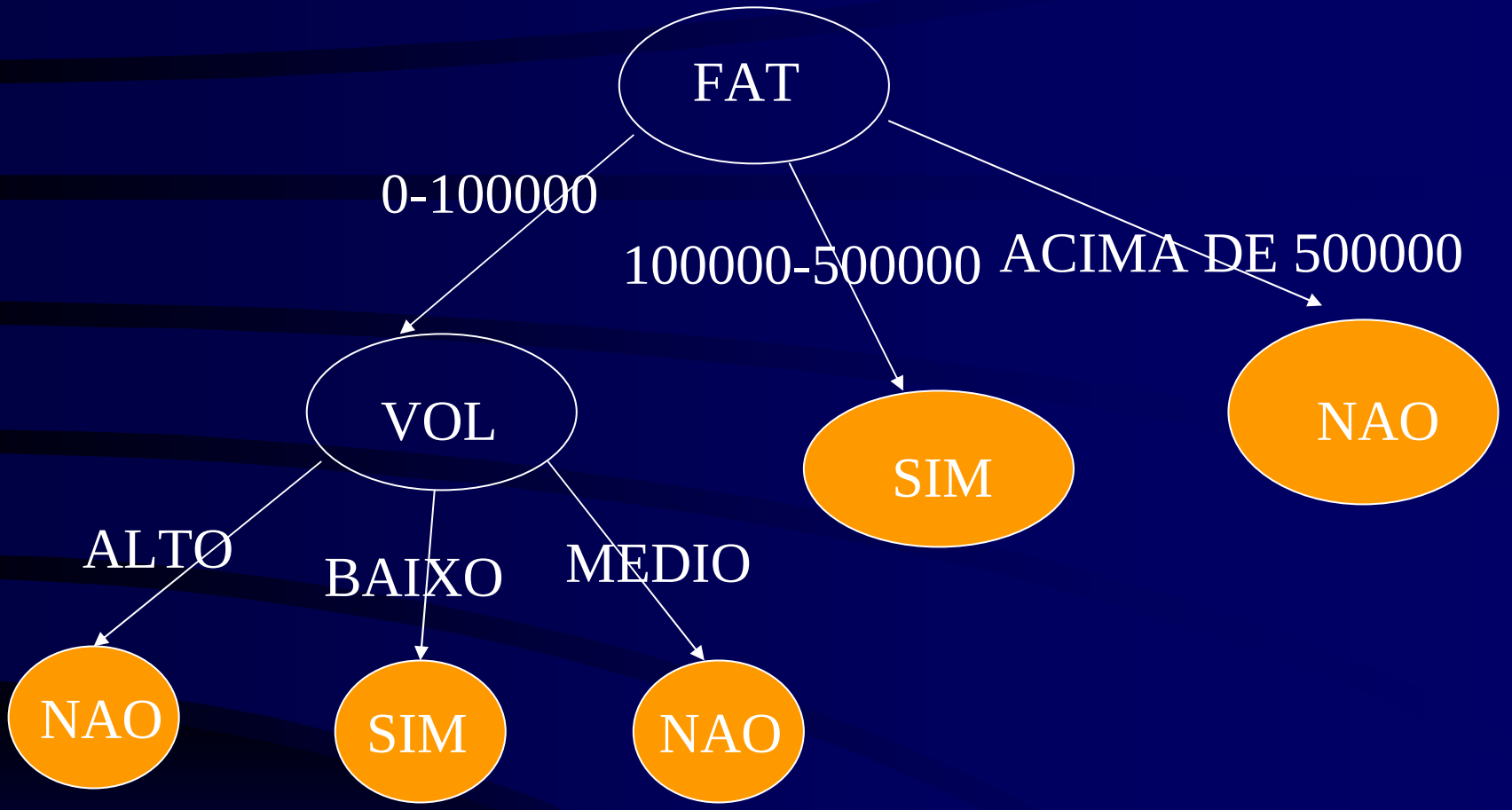
Atributo FAT possui maior valor de Gain,
logo será o atributo inicial ou raiz da árvore.

Processo de formação da árvore



NOME	FAT	VOL	D.VCER	D.CIDA	METAS	LIMCRE	CONJ	SPC	AD
Alberto Reis	100000-500000	MEDIO	NÃO	NÃO	NÃO	ALTO	SIM	SIM	SIM
Castelo Branco	100000-500000	BAIXO	NÃO	NÃO	NÃO	ALTO	SIM	NÃO	SIM
Claudio Tafarel	ACIMA 500000	BAIXO	SIM	NÃO	SIM	BAIXO	SIM	SIM	NAO
Pedro de Assis	ACIMA 500000	MEDIO	SIM	SIM	SIM	BAIXO	SIM	SIM	NAO
Caio de Abreu	0-100000	BAIXO	NÃO	NÃO	NÃO	MEDIO	NAO	NAO	SIM
Jardel de Souza	0-100000	MEDIO	NÃO	SIM	NÃO	ALTO	SIM	SIM	NAO
Jose da Silva	0-100000	ALTO	NÃO	NÃO	NÃO	ALTO	NAO	SIM	NAO

Processo de formação da árvore (cont.)



NOME	FAT	VOL	D.VCER	D.CIDA	METAS	LIMCRE	CONJ	SPC	AD
Caio de Abreu	0-100000	BAIXO	NÃO	NÃO	NÃO	MEDIO	NAO	NAO	SIM
Jardel de Souza	0-100000	MEDIO	NÃO	SIM	NÃO	ALTO	SIM	SIM	NAO
Jose da Silva	0-100000	ALTO	NÃO	NÃO	NÃO	ALTO	NAO	SIM	NAO

Desenvolvimento

- **Desenvolvimento do Protótipo**
  - **– Especificação**
    - **Análise entruturada**
  - **– Banco de dados**
    - **Oracle**
  - **– Ferramentas**
    - **Oracle Forms**
    - **Oracle Graphics**

Desenvolvimento (continuação)

- **Desenvolvimento do SIE**
  - Aquisição dos dados
  - Acesso aos dados

    Utilização do SIE especificado seguindo os processos de KDD:
    - Domínio da Aplicação;
    - Seleção dos Dados;
    - Pré-processamento e limpeza;
    - Data Mining;
    - Interpretação do Conhecimento.

Conclusões

- O Data Mining devolve informações que são induzidas dos dados;
- O Data Mining juntamente com as etapas de KDD se mostrou bastante eficiente para o desenvolvimento do Sistema;
- Testes com o JEVirtual mostraram a eficiência para a construção de modelos;
- Desvantagens do uso de Redes Neurais;
- Os objetivos do trabalho foram atingidos.

Conclusões (continuação)

- Limitações
  - Regras de pré-processamento limitadas;
  - Fonte de dados externa fixa.
- Dificuldades
  - Bibliografia;
  - Componente.

Sugestões

- Aplicação do Data Mining em outras tarefas/técnicas, como Classificação com o uso de Estatística padrão;
- Possibilidade de se escolher entre mais fontes de dados;
- Acesso aos dados através de outros Bancos de Dados.