Departamento de Informática

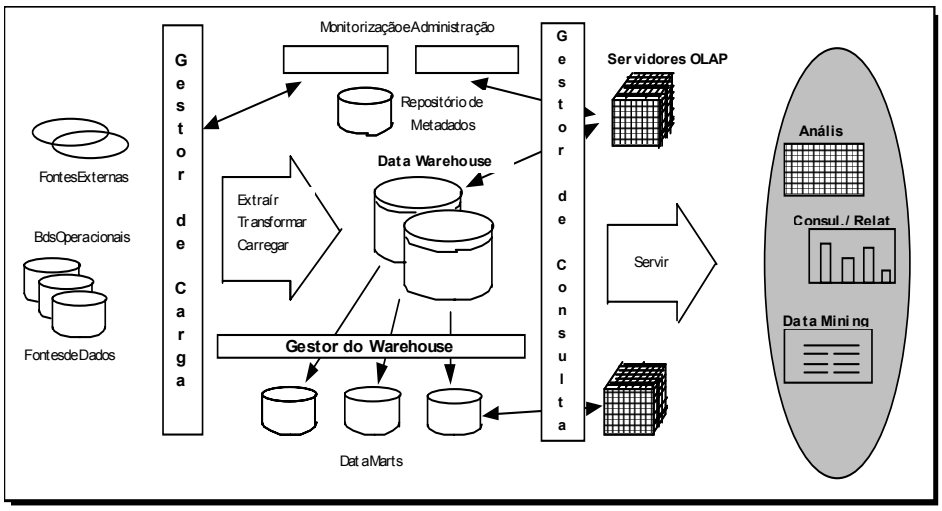
Curso de Engenharia e Sistemas e Informática
5º.ANO
Ano Lectivo 2006/2007 – 1º. Semestre

Teste - Exemplo de Análise Inteligente de Dados (Teórica)

Duração 75 minutos

Obs. A classificação da frequência é obtida através da fórmula: 60% * Freq. Teórica + 40% * Freq. T.Prática

1. (2 V) Na figura ao lado, é mostrado o referencial informacional, mas numa perspectiva funcional. Mostre a relevância de cada componente mostrado no diagrama, esp0ecialmente no que toca à respectiva funcionalidade e interacção.



2. (2 V) Indique e fundamente as diferenças fundamentais entre SQL e OLAP vs. KDD.

3. (2 V) Muitas vezes, quando uma determinada coluna é particularmente importante, efectua-se a chamada amostragem estratificada. Diga em que consiste, porque é importante na situação descrita e dê exemplo.

4. (2 V) As arquitecturas fisicamente bidimensionais, características das tabelas relacionais, seriam, à partida, factores limitativos na utilização de RDBMSs como repositórios para arquitecturas OLAP, o que realmente não acontece.
   Mostre como se resolve esta aparente limitação.

5. (3 V) O Data Mining é possibilitado pela maturidade de quatro tecnologias:
   - armazenamento maciço de dados
   - poderosos computadores multidimensionais
   - algoritmos de data mining
   - visualização de dados

   Diga em que medida é que cada uma destas tecnologias influem decisivamente para o data mining e seu sucesso.

<u>Nas questões seguintes, deverá responder à A ou B.</u>

6. A. (2 V) O facto de o K-NN, não gerar um modelo, constitui uma força e simultaneamente uma fraqueza do método. Comente a afirmação.

6. B. (2 V) Com o K-NN, se K=1, o que tentaremos encontra para efectuar a predição? Discuta a validade dessa abordagem.

7. A. (2 V) A divisão em cada nó da árvore (nos algoritmos de árvores de decisão) é efectuada de forma a que os sub-nós criados sejam mais dissemelhantes uns dos outros e homogéneos em cada um. Mostre que cada divisão constitui um clustering, mas dirigido.

7. B. (2 V) A chamada análise de sensibilidade, permite responder, ainda que de forma indirecta, a uma limitação das redes neuronais. Diga qual e mostre a sua funcionalidade.

8. (2.5 V) Um dos problemas dos algoritmos genéticos surge sob a forma da chamada convergência prematura. Em que consiste, que paralelos se encontram no mundo biológico e em que medida estes últimos forneceram soluções?

9. (2.5 V) A qualidade dos dados internos disponíveis, a sua dispersão e transformações necessárias, além da obtenção de dados externos, é determinante para o sucesso de um projecto de ECBD. Discuta estes tópicos relativamente à problemática de um projecto de ECBD.

Departamento de Informática

Curso de Engenharia e Sistemas e Informática
5º.ANO
Ano Lectivo 2006/2007 – 1º. Semestre

Teste - Exemplo de Análise Inteligente de Dados (Teórico-Prática)

Duração 60 minutos

Um grande operador americano de serviços telefónicos abordou a equipa Darwin (produto de Data Mining Oracle) com o seguinte problema: Seria possível predizer que clientes de serviços locais teriam mais probabilidade de se tornarem clientes lucrativos de longa distância?

A legislação recente de desregulamentação criou simultaneamente uma oportunidade e um risco para a empresa. Agora já não estava proibida de expandir o seu leque de serviços e produtos, mas também tinha perdido a sua posição protegida no mercado. Ao entrar no mercado de serviços de longa distância, a empresa esperava aumentar o valor das suas propostas aos clientes existentes, enquanto alargava o seu mercado.

Estava claro que as contas "lucrativas" não eram simplesmente definidas como aquelas que poderiam ser adquiridas a baixo custo. De facto, os fornecedores de serviços perderam milhões por ano relativamente a clientes com utilização muito baixa - clientes cuja utilização não é suficiente para cobrir mesmo os custos de facturação e despesas administrativas.

Em vez disso, foram definidos como lucrativos - relativamente ao valor do cliente no seu tempo de vida - através de um cálculo onde a propensão do cliente por utilizar serviços de longa distância tinha um peso elevado.

A empresa de telecomunicações já tinha adquirido mais de um milhão de clientes de longa distância no ano anterior. Dessa forma estavam disponíveis montanhas de dados para a criação de modelos. A empresa também tinha já enriquecido os seus próprios dados de marketing e dados operacionais com informação demográfica originária de múltiplas bases de dados comerciais e governamentais.

Uma equipa de projecto compostas por elementos da Darwin e da própria empresa foi criada rapidamente e um conjunto de dados de muitos gigabytes foi carregada no Darwin. O processo de criação de modelos decorreu rapidamente, graças à automatização proporcionada pelos wizards Darwin e pela sua capacidade de processamento paralelo. Realmente, o pessoal da empresa ficou espantada pela velocidade da ferramenta Darwin - um projecto anterior utilizando os mesmos dados levou três meses a completar, utilizando ferramentas desktop!

A prova de qualquer projecto de data mining, contudo, reside na precisão dos seus resultados - e aí Darwin mostrou o seu poder. Apenas numa semana, a equipa do projecto criou um modelo preditivo que ultrapassou todos os exercícios de modelação anteriores.

A equipa desenvolveu recentemente listas ordenadas de prospecção que mostraram um melhoramento de 300% relativamente à selecção aleatória - permitindo à empresa dirigir o seu esforço de marketing às pistas mais rápidas.

Depois de ler atentamente o texto acima que descreve um caso de aplicação de técnicas de data mining, responda às questões seguintes:

1. (3 V) Teria sido uma abordagem exploratória de interesse prático nesta empresa? Fundamente a sua opinião.

2. (3 V) Na sua opinião, a inclusão de pessoal da empresa de telecomunicações no estudo de Data Mining, terá tido que objectivos? Fundamente a sua resposta.

3. (4 V) No texto alude-se à prévia aquisição de mais de um milhão de clientes de serviços de longa distância. Mostre o valor desses clientes para o estudo e que possíveis alternativas haveria se não existissem.

4. (3 V) No texto fala-se também no enriquecimento dos dados operacionais e de marketing internos através de informação demográfica obtida de bases de dados comerciais e governamentais. Mostre o seu interesse conjugado com a existência dos mais de um milhão de clientes de longa distância e no caso destes não existirem de todo.

5. (3 V) A ferramenta de data mining Darwin é extremamente rápida, neste caso numa semana, conseguiu o que outras tinham obtido só ao fim de 3 meses e com uma precisão muito maior. Como será tal possível?

6. (4 V) A melhoria de 300% face à selecção aleatória, mostrada no gráfico, torna-se especialmente interessante, já que é particularmente eficaz para clientes fortes. Comente estas vantagens obtidas pelo modelo gerado pela ferramenta e que sugere seja feito no futuro