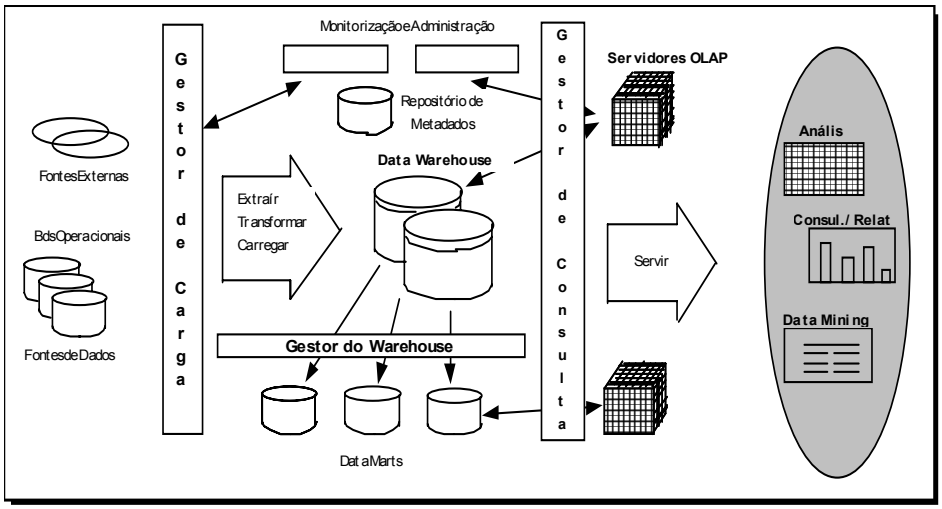
Departamento de Informática

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

Frequência de Análise Inteligente de Dados (Teórica)
Data: 2007/01/18
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 do componente Gestor de Carga, indicando e justificando a sua relação com o componente Gestor de Consultas, metadados, administração e monitorização.



2. (2 V) Distinga predictores contínuos de categóricos, mostrando como efectuar a respectiva conversão, indicando e justificando se é sempre uma operação desejável e mostrando a sua aplicabilidade em redes neuronais.

3. (3 V) O objectivo do processo de KDD é a obtenção de conhecimento útil a partir de grandes colecções de dados. O KDD pode ser visto como um processo integrando várias fases:
   - Compreensão do domínio
   - Preparação do conjunto de dados
   - Descoberta de padrões
   - Pós-processamento dos padrões descobertos
   - Colocar os resultados em uso
     - *Retirado de "Data Mining: machine learning, statistics and databases" Heikki Mannila*

   Discuta cada passo do processo, mostrando a sua relevância e em que medida a realimentação será importante.

4. (2 V) De entre as diversas arquitecturas OLAP, sobressaem as denominadas MOLAP e ROLAP. Distinga-as, efectuando uma análise comparativa entre elas.

5. (2 V) As vertentes predictiva e descritiva são dois dos objectivos do emprego de técnicas de Data Mining. Diga em que consiste uma e outra, como estão relacionadas e em que medida as diferentes técnicas lhe dão suporte.

Nas questões seguintes, deverá responder à A ou B.

6. A. (2 V) Diz-se que o K-NN, não é propriamente uma técnica de aprendizagem, mas mais um método de procura. Dê a sua opinião quanto à validade desta afirmação, fundamentando-a convenientemente.

6. B. (2 V) Com o K-NN, se K=1, o que tentaremos encontrar para efectuar a predição? Discuta a validade dessa abordagem e mostre o paralelismo com a denominada sobreaprendizagem em outras técnicas.

7. A. (2 V) Na indução de árvores de decisão, o problema reside na selecção do preditor a utilizar para a divisão e, em caso de preditores contínuos, qual o valor pelo qual efectuar a divisão. Mostre a validade da afirmação discutindo a forma de como a indução da árvore é efectuada.

7. B. (2 V) Uma característica do algoritmo de divisão de árvore é ser *greedy*, constituindo tal simultaneamente uma vantagem e um óbice. Comente a afirmação, mostrando também como é aliviado esse óbice, nos algoritmos mais recentes.

8. A. (3 V) A vida é, na sua essência, um processo contínuo de busca de melhores soluções. Não será de estranhar assim o facto de muitos algoritmos de optimização e pesquisa de soluções em sistemas complexos sejam inspirados na "vida". De entre eles, podem citar-se as redes neuronais, algoritmos genéticos, optimização por enxame de partículas e sistemas de colónias de formigas. Mostre a veracidade da afirmação, relativamente ao primeiro e segundo dos métodos citados.

8. B. (3 V) Em alternativa aos algoritmos de Gradient-Descent (cujo exemplo estudado foi o de retropropagação), poderemos utilizar algoritmos genéticos na determinação dos pesos, ainda que estes não pareçam obter melhores e mais rápidos resultados do que aqueles. Efectue um comparativo da utilização de uns e outros no processo de aprendizagem das redes neuronais.

9. (2 V) Abaixo é mostrada uma tabela em que a categorização de alguns dos algoritmos de data Mining é efectuada. Discuta o tipo de procura referida relativamente ao 1.º, 3.º e 4.º dos algoritmos enumerados.

-

Algoritmo	Estrutura	Procura	Validação
CART	Árvore Binária	Divisões escolhidas pela entropia ou métrica GINI	Validação Cruzada
CHAID	Árvore Dividida Múltipla	Divisões escolhidas pelo texto do Qui-Quadrado e ajuste Bonferroni	
Redes Neuronais	Rede retropropagada com threshold não linear	Retropropagação dos erros	Não aplicável
Algoritmos Genéticos	Não aplicável *	Sobrevivência do mais apto em mutação e cruzamento genético	Normal/ a validação cruzada
Indução de Regras	Regra If ... then	Adiciona novos constrangimentos às regras e reem-nas se passar no critério de interesse - precisão, cobertura, etc.	teste do Qui-quadrado com significância estratística
|Nearest Neighbor	Distância do protótipo no espaço n-dimensional	Normalmente não há procura	Validação cruzada utilizada para teste da taxa de precisão

Departamento de Informática

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

Frequência de Análise Inteligente de Dados (Teórico-Prática)
Data: 2007/01/18
Duração 60 minutos

CASO 1

Lowestfare.com necessitava de conhecer melhor os seus clientes por forma a saber quais seriam passíveis de utilizar o canal de vendas Internet. Seleccionando os campos mais valiosos de dados a adquirir externamente para popular o Data Warehouse e através da identificação dos clientes com maior apetência de compras através da Internet, Oracle Data Mining forneceu inteligência para o negócio que resultou em poupanças significativas para a empresa.

Lowestfare.com é um fornecedor de produtos e serviços relacionados com viagens de lazer e negócios. Vende bilhetes de avião, reservas de hotéis, aluguer de carros, cruzeiros e pacotes de viagens. Estas vendas são realizadas através de três canais diferentes: Internet, *call centers* e agências de viagens.

A empresa iniciou as suas operações em 1995 com a venda de bilhetes, acrescentando, sucessivamente, novos serviços. Compreendeu agora que, por forma a manter-se competitiva, deverá tornar-se uma "*One Stop Shop*" de viagens. O problema de negócio que enfrenta, posto de uma forma simples, é a necessidade de um melhor conhecimento dos seus clientes por forma a poder vender-lhes os produtos certos, através do melhores canais. Fazendo isto, aumentar-se-á a lealdade dos clientes, o que terá como consequência, a longo prazo, o aumento dos proveitos e lucros.

A primeira coisa que a empresa pretende, relativa aos seus clientes, é saber quais terão maior apetência de compras através da Internet. O conhecimento deste clientes - alvo permitir-lhe-á aumentar o lucro por cada bilhete vendido. A longo prazo, o melhor conhecimento de quem são os seus clientes e quais as suas necessidades darão à empresa oportunidade de aumentar a lealdade dos seus clientes e rendibilidade respectiva.

Mas há um óbice imediato: a empresa dispõe de muito pouca informação de carácter não restritamente operacional, relativa aos seus clientes. A aquisição de bilhetes por parte de clientes não obriga ao fornecimento de qualquer informação de carácter demográfico. Para um melhor conhecimento os seus clientes teriam de adquirir informação demográfica e adicioná-la à informação que dispunham já sobre os seus clientes.

Depois de uma procura exaustiva, escolheram Acxiom como tendo a informação de que necessitavam. Esta empresa dispunha de uma enorme quantidade de informação que poderia ser adicionada a cada registo de cliente, mas que seria tanto mais cara, quanto mais informação fosse adquirida. Eventualmente, nem todos, ou mesmo só uma pequena parte (dos cerca de 650 atributos disponíveis), teria algum valor para a construção de modelos ou extracção de outro tipo de conhecimento. Assim, adquiriu-se a totalidade da informação disponível, mas relativa a só alguns meses, e empreendeu-se uma análise exploratória, para verificação da relevância de cada um dos campos. Esta análise exploratória constituiu o primeiro passo do projecto. Daqui resultou a selecção de 87 campos que se mostraram relevantes para as necessidades futuras de negócio, evitando os custos que seriam suportados com a aquisição da totalidade dos campos irrelevantes para o negócio.

O passo seguinte foi a utilização do Oracle Data Mining Suite para ajudar a compreender melhor o perfil dos clientes, baseados nas compras pela Internet, *call centers* e agências de viagens. Foram desenvolvidos cerca de 30 modelos exploratórios e 20 modelos para desenvolvimento que formam depois convertidos num modelo de produção. Este foi depois utilizado para estabelecer um *ranking* dos cientes da Lowestfare.com com mais apetência para aquisições pela Internet, contribuindo para uma enorme redução de custos.

1. (4 V) Enquadre as necessidades sentidas pela empresa, numa perspectiva de análise de dados e mostre a necessidade sentida de empreender o enriquecimento dos dados disponíveis internamente.

2. (3 V) Os modelos criados foram sendo provavelmente objecto de melhorias sucessivas. Indique, justificando, algumas das actuações, através das quais poderão ter sido conseguidas e como terão sido sucessivamente avaliados.

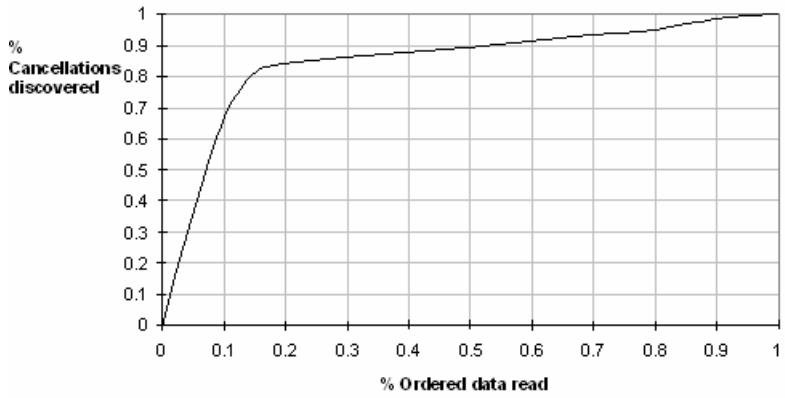
3. (4 V) Numa perspectiva de aquisição de novos clientes, mostre como poderia utilizar na prática o modelo ora criado, como poderia ser melhorado e, inclusivamente, ser de utilização em tempo real (no canal Internet).

CASO 2

Seguros Winterthur – precisão na predição da lealdade dos clientes

Os Seguros Winterthur podem agora predizer quais os clientes que irão cancelar as suas apólices com 90% de precisão. Isto é o resultado de uma avaliação comparativa de produtos, na qual empresas produtoras de produtos de mineração de dados foram convidados a produzir modelos preditivos utilizando um grande conjunto de dados. Os modelos foram testados pela Winterthur em dados reais. Dos vários produtos testados, o Clementine produziu o melhor modelo.
Winterthur tem mais de um milhão de clientes em Espanha, onde a avaliação foi levada a cabo – e mais de 130,000 cancelam as suas apólices em cada ano. Dadas as perdas de rendimentos e o custos de obtenção de novos clientes, trata-se claramente de um problema "caro".

a) O conjunto de dados teste inicial era relativo a apólices de seguros no ramo automóvel, com registos contendo 250 campos que descrevem cada caso. Utilizando conhecimento de negócio e técnicas de visualização de dados, o número foi reduzido. Várias técnicas foram utilizadas para ajudar a seleccionar os campos mais significativos.



The data set was ordered according to propensity to cancel, predicted by Clementine. As a result, most of the cancellations now come near the beginning of the data

b) O modelo completo desenvolvido consistiu numa combinação de redes neuronais. Ele predisse correctamente quem cancelaria a sua apólice com 90% de precisão num conjunto de dados teste para avaliação dos modelos.

c) A cada cliente foi dado um "score", que indica a apetência para cancelar apólices. Os dados foram então ordenados pelo "score" e verificou-se que 75% dos clientes a perder potencialmente, surgiam nos primeiros 15% do conjunto de teste, como é mostrado no gráfico acima. Este tipo de "ranking" é uma ajuda valiosa nos esforços de focalização para retenção de clientes.

4. (3 V) Mostre a necessidade focada em a) e como poderá ter sido conseguida.
5. (3 V) Em b) refere-se que as redes neuronais foram a forma eleita para os modelos. Justifique a opção e sugira outras alternativas, fundamentando-as.
6. (3 V) Qual a relevância da constatação descrita em c) e como pode ser utilizada?