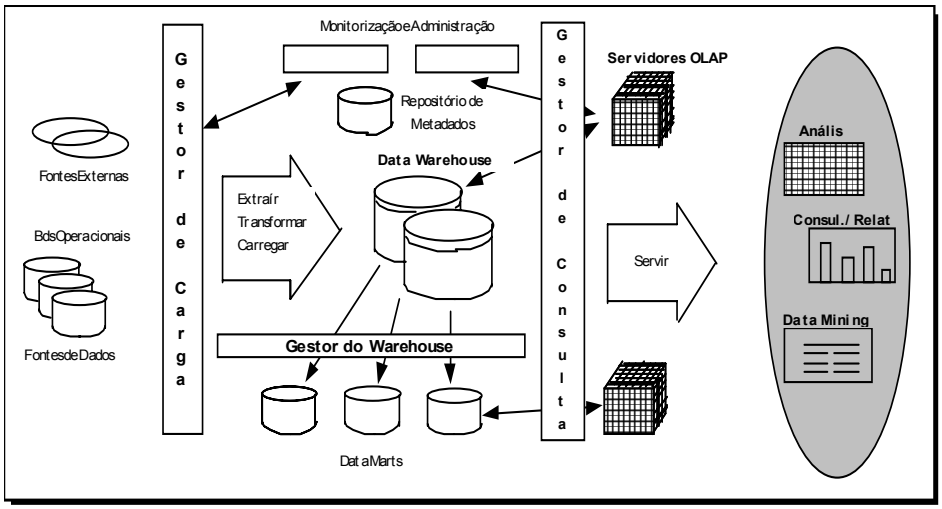
Departamento de Informática

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.



Dicas: O gestor de carga é responsável pelo processo de ETL, utilizando os metadados para gerir o processo; as alterações de perfil detectadas e reportadas pelo gestor de consultas ao componente de administração e monitorização e podem implicar a alteração das agregações ou mesmo, implicar, por decisão do administrador do DW, a alteração do próprio esquema do DW e DMs e respectivo processo de ETL. Estas alterações serão reflectidas nos metadados. Estes terão, assim, uma função dual: podem ser utilizados para gerir o processo de ETL, a alteração dos mecanismos de ETL (quando necessário) e do lado das consultas, servindo para mostrar os dados disponíveis, e escondendo as complexidades da própria base de dados.

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.

Dica: As duas primeiras questões podem ser resolvidas facilmente com a consulta aos diapositivos da disciplina. Quanto à necessidade da operação, prende-se com o tipo de algoritmo a utilizar. Para o caso das redes neuronais, como tratam apenas preditores contínuos, trata-se de uma operação não absolutamente necessária, ainda que possa também ser utilizada (ver quando). Antes é necessário um esquema para poder utilizar preditores categóricos (o problema inverso). Para o fazer ver nos diapositivos como fazer a preparação de dados em variáveis categóricas e perceber realmente porquê.

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.

Dica: Neste caso, apenas a questão da realimentação pode levantar dúvidas. A realimentação vai permitir uma avaliação real do processo KDD utilizado. Esta pode implicar, por uma lado, alterações ou recriação de todo o processo; por outro lado, como a utilização dos resultados vai originar novos dados, que podem servir para gerar novo modelo, a realimentação vai permitir efectuar um ajuste dinâmico do processo.

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

Dica: A consulta dos acetatos e de uma das fichas práticas será suficiente para responder a esta questão.

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.

Dica: A vertente preditiva está relacionada com predição, predizer algo baseado num conjunto de dados original, materialização do comportamento de uma qualquer entidade ou processo, a partir do qual pode ser criado um modelo ou pode ser utilizado directamente, para pesquisa de casos idênticos. Apresentado um novo registo, é gerada a respectiva predição. Um exemplo pode ser o caso da predição relativa aos casos de empréstimos bancários: gerado o modelo, utilizando os dados correspondentes a casos de empréstimos anteriores (necessariamente incluindo clientes cumpridores e não cumpridores) é possível, apresentado um novo cliente possível, avaliar o respectivo risco. Em resumo, o que se pretende é prever algo para um novo caso.
Quanto à vertente descritiva pretende mostrar uma visão de alto nível sobre a entidade ou processo de que os dados são mostra do comportamento. Pretende-se compreender o próprio comportamento do processo ou entidade de uma forma geral, bastante sucinta.
Quanto ao suporte das diversas técnicas, pode ser percebido por consulta aos diapositivos, sendo paradigmático o caso das redes neuronais MLP e árvores de decisão, mas será conveniente incluir outros.

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

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.

Dica: No K-NN não há geração de modelo e é uma técnica baseada em distância, por procura de registos mais próximos.

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.

Dica: Se K=1 estaremos a procurar um único caso, o mais idêntico possível ao novo caso. Realmente, estamos a procurar eventualmente casos particulares, não generalizáveis e assim, pouco passíveis de revelar um comportamento de alto nível escondido nos dados utilizados. O

mesmo se passa quando há sobreaprendizagem por exemplo na árvores de decisão: o modelo vai reflectir o nível de detalhe – registo – não sendo, de forma alguma, geral.

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.

Dica: A indução das árvores de decisão é feita sob dois pressupostos fundamentais: 1) todos os registos num nó pai, estarão contidos nos nós filhos gerados pela divisão; 2) os nós filhos devem ser o mais homogéneos em si e mais heterogéneos entre eles. Assim, este último pressuposto implica que a divisão deve ser efectuada de forma a que a diversidade nos nós gerados seja a mínima possível, ou seja, que a entropia seja mínima (haja uma diminuição máxima de entropia). Assim, para que este objectivo seja conseguido, é necessário seleccionar o preditor que, ao ser utilizado como discriminante (usados para efectuara a divisão) o vai permitir. Claro que se o preditor for contínuo ou tiver vários valores possíveis, importa também seleccionar qual o valor que vai assegurar a divisão de diversidade mínima.
Quanto à forma, basta mostrar um desenho com um caso e árvore simples (por exemplo, aquela relativa ao caso dos empréstimos) e mostrar como é efectuada para um dado nó.

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.

Dica: Nos diapositivos encontra-se a resposta a esta questão, pelo que não serão dadas quaisquer explicações adicionais.

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.

Dica: Rede neuronal - A comparação entre a arquitectura de uma rede neuronal (várias camadas, ligações e utilização de elementos tipo perceptrão) e os muitos neurónios do cérebro, com os dendritos e axónios como ligações, gerando a arquitectura respectiva. Quanto a algoritmo genético, emula os sistemas biológicos: populações de indivíduos (cada um com um genoma), cruzamentos, mutações e selecção natural, por avaliação da adaptação de cada um dos indivíduos. Pode utilizar-se depois um esquema para mostrar como funciona um algoritmo genético.

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.

Dica: Quanto ao gradient-descent, vem descrito nos diapositivos, onde poderá ser consultado. Já a utilização de um algoritmo genético, e uma vez que é necessário saber como codificar o problema e como avaliar a solução, vai aqui tratar-se o assunto. Quanto à primeira questão, basta que o genoma de cada indivíduo inclua um valor para cada um dos pesos (ligações entre os perceptrões nas várias camadas que constituirão a rede. Já a segunda questão, a avaliação (qualidade do modelo) será relativa normalmente à precisão do modelo. Como cada indivíduo será uma solução para o modelo da rede neuronal, bastará aplicar os casos e ver, para cada um, qual será a adaptação de cada um (avaliada, como se disse, em relação à veracidade da predição efectuada). A

evolução (as sucessivas gerações, com aplicação dos operadores genéticos) encarrega-se da procura da melhor solução.

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

Dica: As respostas podem ser procuradas nos diapositivos, de forma directa, pelo que não serão dadas quaisquer outras explicações.

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.

Dica: Iniciar com um enquadramento cerca da relevância da Internet como canal de contacto e vendas. Falar depois sobre a necessidade sentida pela empresa de se tornar numa "*One Stop Shop*", o que implica falar sobre a necessidade de conhecer a apetência dos clientes pelo canal Internet, já que de cada vez maior relevância. Mas, a empresa não dispõe dos dados necessários, já que apenas dados de valor operacional tem sido recolhido. Assim, importa, porventura, empreender uma abordagem dual: 1) Para resposta imediata e directa ao problema em mãos, importa adquirir dados que permitam colmatar a falha detectada. Há um custo financeiro a suportar, mas, por uma lado, pode ser limitado (pela aquisição de apenas os dados que realmente interessam), e, por outro lado, espera-se, com grande grau de certeza, que o estudo permita um ROI elevado, como é apanágio da maioria das iniciativas de BI correctamente empreendidas. 2) para o futuro, importa alterar o próprio sistema operacional, de forma a promover a recolha dos dados necessários (relativos a novos clientes ou para completar a informação relativa a clientes actuais), dados estes que deverão depois ser objecto da devida purificação.

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.

Dica: Uma das estratégias possíveis a ser seguida consiste na utilização de várias técnicas. No texto não é dito que tipo de análise foi empreendida, mas talvez a segmentação dos clientes, extracção de regras de associação e predição, poderão ter sido empreendidas. Assim, iniciar com uma análise estatística e do emprego de técnicas de visualização que permitem a obtenção de conhecimento acerca de uma visão global sobre os dados e de dicas acerca do comportamento global do fenómeno que se está a estudar, importa depois empreender a aplicação de técnicas de mineração de dados. Quanto à segmentação pode utilizar-se as várias técnicas: clustering não hierárquico e redes Kohonen. O clustering hierárquico poderá ser utilizado também, apesar do seu elevado custo computacional, mas importa conhecer o número de clusters que os próprios dados irão induzir, apesar de um conjunto de tentativas sucessivas p. ex. com as redes Kohonen, possa também conduzir ao mesmo resultado. Para predição, começar com algoritmos de aprendizagem rápida, interessantes para uma análise exploratória (ex. Naive-Bayes), e depois utilizar redes neuronais e algoritmos para indução de árvores de decisão. A avaliação aqui pode ser efectuada por comparação da precisão e compreensão dos modelos gerados. Para regras de associação, utilizar uma (ou duas) técnicas disponíveis na ferramenta utilizada, aí, decerto implicando um maior grau de interacção dos analistas, já que importa sempre a definição da precisão e suporte e avaliar do interesse das regras geradas. Não esquecer que, em cada nova análise, muitas das fases do processo de KDD deverão ser sucessivamente repetidas, especialmente no que toca à preparação dos dados.

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).

Dicas: O modelo gerado pode ser aplicada a novos clientes, desde que disponíveis o conjunto de dados necessários ao novo cliente. P. ex. se o objectivo for o conhecimento da apetência (S/N) pelo canal Internet, teremos um conjunto de variáveis preditoras (de entrada ou independentes) e uma variável de predição (de saída ou dependente). Sabidas os valores das variáveis de entrada para o novo cliente, o modelo (rede neuronal ou árvore de decisão) gerará a resposta. A mesma abordagem poderá ser utilizada para conhecer o segmento a que o novo cliente irá pertencer. Mas isto pode depois ser complementado: depois da oferta para aquisição do cliente, este, ao entrar no site para realizar a compra, deverá ser motivado (conduzido), na fase de registo, a fornecer um

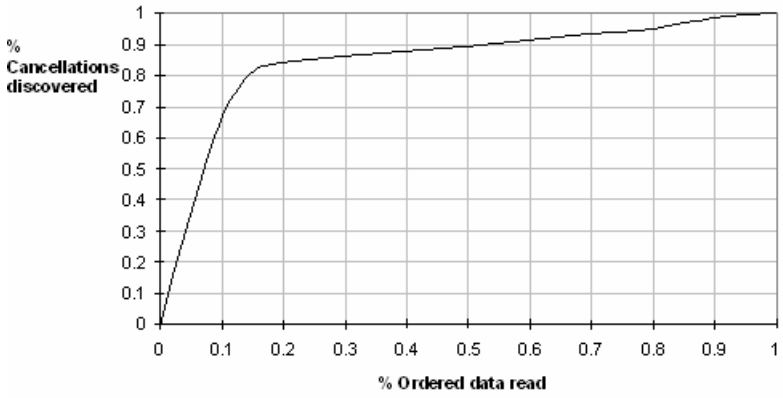
conjunto de dados de valor informacional (processo obviamente repetido para qualquer novo cliente, por maioria de razão). Com estes dados, o sistema vai avaliar o cliente, extrair o seu perfil e inseri-lo num dado cluster. Isto permitirá, sugestões de produtos em tempo real. Também, cada nova aquisição e mesmo, cada manifestação de interesse do cliente, deverá ser avaliada em tempo real, conduzindo à reavaliação do perfil e assim, importar numa actualização dinâmica do perfil que, por sua vez, importará em novas propostas de artigos para venda (com eventuais promoções, se considerado relevante). Na prática, este esquema é algo que, p. ex., é realizado no site da Amazon.

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.
Dica: O número de campos a utilizar deve ser um compromisso entre precisão e desempenho: a utilização de campos não relevantes para a análise em mãos, irá prejudicar necessariamente o desempenho, podendo, inclusivamente, ter efeitos perniciosos na precisão (ao gerar modelos mais complexos, menos adequados à realidade do problema). A compreensão do negócio (a interacção com um especialista) permite inferir quais os campos que, à partida, terão algum valor para a solução de um determinado problema. Técnicas de visualização, estatísticas e de análise de sensibilidade poderão ser utilizadas também, por um lado, para aquilatar da validade das selecções feitas e, por outro, para perceber realmente a força das ligações entre os campos e, assim, tomar a decisão mais acertada. Também, p. ex., a indução de árvores, permite avaliar quais os campos com maior valor discriminatório e, assim, permite uma selecção daqueles mais úteis, algo que pode ser efectuado com um conjunto de dados reduzido (incluindo a totalidade dos campos, mas com

apenas um apequena parte dos registos disponíveis, mas necessariamente uma amostra significativa).

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.

Dica: Focar nas vantagens das redes neuronais MLP (precisão e eficiência na fase de predição). Assim, dado o número previsivelmente elevado de possíveis novos clientes, será uma vantagem a levar em conta. Quanto à questão da precisão, em termos intrínsecos, não carece de mais justificações. Também, e como o número de cancelamentos é da ordem dos 13%, importa utilizar um modelo que consiga uma precisão bastante elevada, já que, a probabilidade à priori de não cancelamento será da ordem dos 67%. Mas, mesmo para os cancelamentos, é-nos dito que o modelo gerado apresenta uma precisão de 90%, o que é óptimo e claramente maior do que a probabilidade à priori (neste caso 13%).

Outras escolhas possíveis serão as árvores de decisão e Naïve-Bayes (descrever as vantagens e desvantagens de cada um) e referir que deverão conseguir precisões bastante razoáveis.

6. (3 V) Qual a relevância da constatação descrita em c) e como pode ser utilizada?

Dica: Como 75% dos potenciais clientes em risco, estão concentrados em 15% dos clientes, isto será uma ajuda valiosa na focagem dos esforços de retenção desses clientes. Por um lado, permite verificar o porquê da perda dos clientes, por análise de apenas uma pequena parte deles. Por outro lado, os esforços de retenção são orientados para a parte realmente em risco, evitando desperdício de recursos.