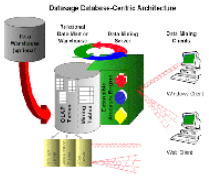
Robert Groth

Métodos e Algoritmos de Data Mining(parte 2)

Usama Fayyad et al



Métodos e Algoritmos de Data Mining

- **Soluções distância (K-NN e clustering)**
- **Naïve-Bayes**
- ⇨ **Arvores de decisão**
- ⇨ **Regras de associação**
- **Redes neuronais**
- **Algoritmos genéticos.**
- **Combinação de múltiplos métodos de predição.**

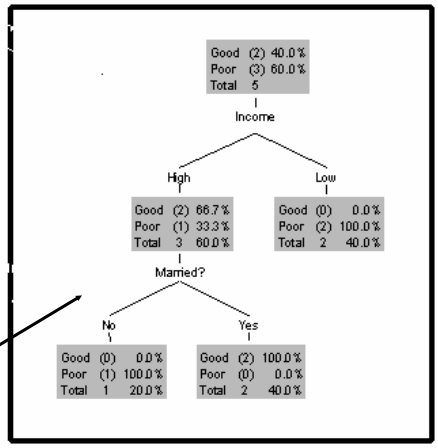
- **Alguns prós e contras das tecnologias mais comuns; ferramentas mais relevantes e suas características**

Árvores de Decisão

Árvores de Decisão

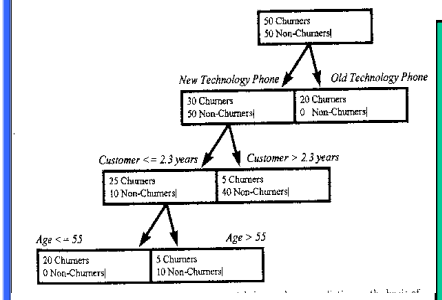
Trata-se de um modelo que é simultaneamente preditivo e descritivo.

- **o seu nome deriva do facto do modelo resultante ser apresentado na forma de uma estrutura em árvore;**
- **cada ramo da árvore traduz uma questão de classificação;**
- **as folhas da árvore são partições dos dados com a sua classificação.**



Aspectos descritivos:
•o débito parece não ter qualquer papel na determinação do risco
•as pessoas com baixo rendimento são sempre de risco pobre
•o rendimento é o factor mais significativo na determinação do risco

Árvores de Decisão



Outro exemplo de árvore de decisão, relativa a exemplo de classificação de clientes que não renovam os contratos relativos a telemóveis (ditos "churn")

Observações:

- **divide os dados em cada nó, sem perder qualquer dado (o nº de registos total num dado nó, é igual à soma dos registos contidos nas suas duas folhas)**
- **o n.º de clientes que não renovaram e renovaram o contrato é conservado, à medida que se sobre ou desce na árvore**
- **é fácil compreender como o modelo está a ser criado (em contraste com os modelos das redes neuronais ou estatística standard)**
- **é fácil utilizar este modelo ao pretender-se atingir os clientes em vias de não renovar o contrato, com ofertas de marketing**
- **pode criar algumas intuições acerca da base de dados de clientes:**

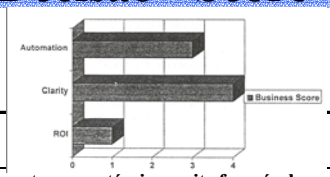
Ex. "clientes há vários anos e que tenham telemóveis actuais, são muito leais"

Árvores de Decisão - generalidades

A apresentação visual torna o modelo de muito fácil compreensão e assimilação, o que motivou que a técnica se tenha tornado muito popular:

- **são normalmente utilizados para classificação (predizer a que grupo pertence um caso);**
- **podem ser também utilizadas para regressão (predizer um valor específico);**
- **de uma perspectiva de negócio, as árvores de decisão podem ser entendidas como uma segmentação dos dados originais (cada segmento será uma das folhas da árvore);**
- **a segmentação é feita por uma determinada razão - para predição de algum pedaço importante de informação. Os registos que caiam dentro de cada segmento, estão lá porque têm semelhança com respeito à informação predita - não só porque são semelhantes (sem que a semelhança seja bem definida).**

Árvores de Decisão - Score



Medida	Descrição
Automatização	As árvores de decisão apresentam uma técnica muito favorável para automatizar a maioria do processo de data mining e modelação predictiva. As suas soluções automatizadas embebidas para coisas como prevenção de sobreadaptação e manuseamento de dados em falta que, em muitas outras técnicas, sobrecarregam o utilizador.
Clareza	O modelo criado pelas árvore de decisão pode ser facilmente visualizado como uma árvore de decisão simples baseadas em predictores familiares ou como um conjunto de regras. O utilizador pode confirmar a árvore de decisão à mão ou modificá-la e dirigi-la, com base na sua própria experiência.
ROI	dado que as árvores de decisão trabalham bem com bases de dados relacionais, traduzindo modelos de árvores de decisão em SQL para tratamento da base de dados, proporcionam soluções bem integradas, com modelos muito precisos.

Resumo da Árvores de Decisão

- O output primário do algoritmo de árvore de decisão é a própria árvore
- O processo de treino é denominado de indução - requer um número pequeno de passagens (normalmente menos de 100)
- Eficiência:

  Naïve-Bayes > Árvores de Decisão > Redes Neuronais
- É um método muito popular, dada a facilidade de compreensão de resultados
- Muitos produtos também traduzem a árvore para regras tipo:

  Se Rendimento = Alto e Anos de trabalho > 5 Então Risco Crédito = Bom
  - De facto os algoritmos de árvores de decisão são muito semelhantes aos algoritmos de regras de indução, que produzem conjuntos sem uma árvore de decisão

História das Árvores de Decisão

- **Resulta da evolução das técnicas surgidas durante o desenvolvimento de disciplinas de machine learning:**
  - **cresceram a partir de uma abordagem analítica chamada de AID (Automatic Interaction Detection), desenvolvida na univ. de Michigan.**
- **AID - trabalha através de teste automático de todos os valores nos dados p/ identificar quais estão mais fortemente associados com o item de saída escolhido para análise. Os valores encontrados que tenham uma associação forte, serão os predictores chave ou factos explanatórios, normalmente chamados de regras.**
- **CHAID - ChiSquared + AID, sendo desenvolvido a partir deste, expandindo as suas capacidades através da adição da fórmula estatística do qui-quadrado.**

História das Árvores de Decisão

- **Na Austrália surge a tecnologia que permitiu o grande crescimento das árvores de decisão - muitas pessoas consideram o prof. Ross Quinland da Univ. Sidney o "pai" das árvores de decisão.**
- **Contribuição - família de algoritmos:**
  - **ID3, 1983 e suas evoluções ID4, ID6, C4.5 e C5.0 que produzem regras relacionadas pelas sua importância, gerando um modelo de árvore de decisão dos factos que afectam o item de saída**
- **Novos algoritmos de árvore de decisão, como GINI, inventado por Ron Bryman, com bom desempenho e capacidades alargadas de processar números e texto.**

Indução da Árvore

A maioria dos algoritmos de árvores de decisão operam em duas fases:

- crescimento da árvore - processo iterativo que envolve a divisão dos dados em subconjuntos progressivamente menores;
- poda (pruning) - técnica que permite tornar a árvore mais genérica, evitando a sobreadaptação.
  - se a árvore cresce até ao máximo, reflectirá todos os detalhes da base de dados, o que, além de tornar a árvore mais complexa, torna-a menos precisa, já que dará como predição, muitas vezes, o resultado de um caso específico.

Depois da indução da árvore, há ainda o necessário teste, para validar o modelo, utilizando um data set independente.

- só depois de determinada a precisão e considerada aceitável. a árvore (ou as suas regras) está pronta a ser utilizada como predictor.

Indução da Árvore - Crescimento

Crescimento da Árvore ocorre por divisão dos ramos:

- os dados são divididos em subconjuntos progressivamente menores;
- cada iteração considera os dados num só nó;
- a 1ª iteração considera o nó raiz, que conterá todos os dados;
- iterações subsequentes trabalham na derivação dos nós que conterão subconjuntos dos dados.

Ao algoritmo colocam-se duas questões:

- que variável independente utilizar para a divisão;
- que valor considerar para a divisão.

Indução da Árvore - Crescimento

O algoritmo começa por analisar os dados para encontar a variável independente (como rendimento, estado civil, débito) que, quando utilizada com uma regra de divisão, resulta em nós que sejam mais diferentes uns dos outros, em relação à variável dependente (valor de predição) e homogéneos em cada um.

- **mede a alteração das densidades dos valores de predição depois da divisão ser feita,**
- **tenta a redução da desordem do segmento original, dividindo-o em segmentos mais pequenos, mais ordenados (mais concentrados em valores particulares de predição).**

As medidas a utilizar para avaliar a diferença, podem ser diversas: algumas implementações utilizam apenas uma, outras deixam o utilizador escolher a medida a utilizar. Ex. entropia, rácio de ganho de entropia, gini e qui-quadrado.

- **Independentemente da medida utilizada, <u>todos os métodos requerem uma tabulação cruzada</u> entre a variável dependente e cada uma das variáveis independentes.**

Característica Importante do Algoritmo de Divisão

O algoritmo de divisão da árvore é Greedy:

- **permite assim que as decisões seja locais e não globais.**

Vantagem:

- **permite reduzir enormemente a complexidade e tempo de indução;**

Desvantagem:

- **A análise de decisão não é global e, assim:**
  - **quando se decide da divisão num determinado nó, um algoritmo greedy não olha para trás na árvore para ver se outra decisão num nó anterior (avaliada em conjunto com a decisão do nó actual) produziria um melhor resultado global.**

Ou seja, <u>poderia haver uma divisão anterior</u> (mais próxima da raiz) que não era melhor (se avaliada apenas localmente), mas <u>que, se tivesse sido utilizada, resultaria numa árvore com uma melhor precisão global</u> (avaliada em conjunto com a possível divisão agora em consideração)

Exemplo Tabulação Cruzada

Voltando ao set relativo ao caso de empréstimo da instituição de crédito, criando a tabela cruzada, teremos:

Nome	Débito	Rendim.	Casado?	Risco
Joaquim	Alto	Alto	Sim	Bom
Susana	Baixo	Alto	Sim	Bom
João	Baixo	Alto	Não	Pobre
Maria	Alto	Baixo	Sim	Pobre
Frederico	Baixo	Baixo	Sim	Pobre

temos de avaliar qual será o predictor com maior influência no resultado;

neste caso, bastará tomar aquele que tiver o maior número de instâncias na sua diagonal principal.

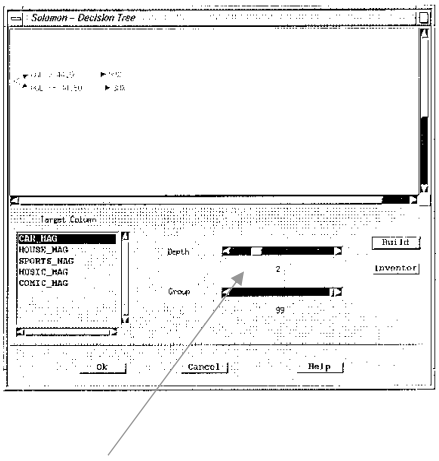
Risco	Débito Alto	Débito Baixo	Rendim. Alto	Rendim. Baixo	Casado	Solteiro
Bom	1	1	2	0	2	0
Pobre	1	2	1	2	2	1

Esta medida é menos ad hoc do que parece, pois que vai ser escolhida a divisão que tenha um menor n.º de instâncias que se desviam do valor predominante da variável dependente em cada nó formado pela divisão.

neste caso será o rendimento (4 instâncias)

Indução da Árvore - Quando Parar o Crescimento

- **Nós puros (que contenham apenas instâncias com o mesmo valor da variável dependente), não serão mais divididas;**
- **Outras regras de paragem, baseadas em:**
  - **profundidade máxima da árvore;**
  - **número mínimo de elementos num nó considerado para divisão;**
- **Em muitas implementações o utilizador pode alterar estes parâmetros.**



Profundidade Máxima da Árvore

Porque não criar a árvore até à sua profundidade máxima?

- **Teríamos uma árvore pura, desde que não houvesse registos em conflito nos dados de treino**

registos em conflito - serão aqueles que têm os mesmos valores para as colunas independentes, mas valores diferentes para a coluna dependente. Não há maneira de criar uma divisão para os diferenciar, a não ser que nova coluna seja introduzida

Mas uma árvore com profundidade máxima será desejável?

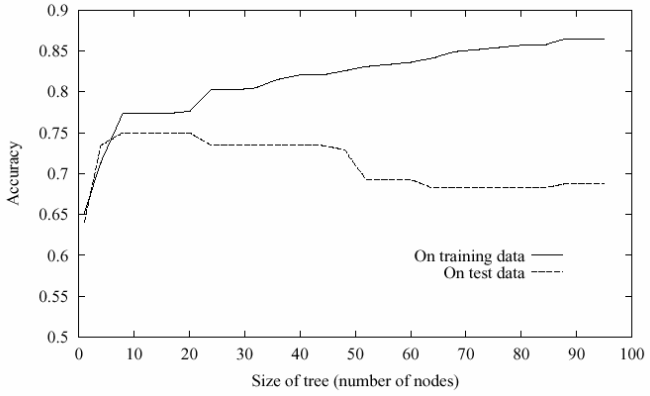
- **alguns algoritmos começam por criar este tipo de árvore (precisão máxima para o conjunto de treino)**
- **mas essa árvore é, em regra, sobreadaptada (overfit)**

uma arvore overfit (sobreadaptada) pode ser vista em termos de predição como tentando encontar um registo do set de treino que seja em tudo semelhante ao novo registo, predizendo o valor da variável dependente, baseado no valor encontrado nesse registo
Problema: uma árvore com profundidade máxima é uma árvore demasiado específica, não encontrando os princípios gerais que estejam subjacentes nos dados.

Assim:

- **ou se limita o crescimento para evitar a sobreadaptação**
- **ou se executa a poda (pruning)**

Sobreadaptação (Overfit) da Árvore



À medida que são adicionados novos nós (com o ID3), no crescimento de uma árvore de decisão, a precisão da árvore, medida relativamente aos exemplos do set de treino, cresce monotonicamente. Contudo, quando medida relativamente a um conjunto de teste, independente dos dados de treino, a precisão cresce de início, depois desce.

Poda (pruning) da Árvore

Operação suplementar de simplificação, que, como se fala em árvores, por analogia, se chama de pruning (poda).
Utilizada para tornar a árvore mais genérica.
Necessária porque:

- **a árvore pode revelar nós ou subárvores que sejam indesejáveis por causa da sobreadaptação;**
- **pode conter regras que o perito no domínio sente serem desapropriadas.**

Pode ser:

- **controlada por parâmetros especificados pelo utilizador;**
  - **ex. a diferença de precisão calculada entre os nós resultantes e o original seja insignificante e abaixo de um determinado limitar;**
- **invocada automaticamente, em algoritmos que induzem árvores com profundidade máxima;**
- **poda interactiva.**

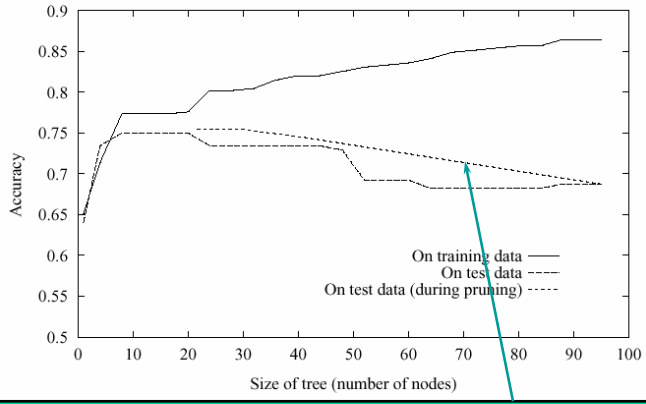
Funcionamento do Algoritmo de Poda

A poda é normalmente boa ideia, mas
como determina o algoritmo onde podar a árvore?

Há várias formas, mas a mais apelativa será:

- **Utilizar uma amostra de controlo com relacionamentos conhecidos e verificados entre as variáveis independentes e dependente;**
- **Comparando o desempenho de cada nó (medido pela sua precisão) relativo às subárvores, torna-se óbvio que divisões têm de ser podadas para atingir a precisão geral mais elevada.**

Efeito do Pruning na Redução de Erros



Efeito do pruning na precisão da árvore de decisão produzida com o algoritmo ID3 (mostrada anteriormente). Note-se o aumento em precisão relativamente ao conjunto de treino à medida que nós são podados na árvore.

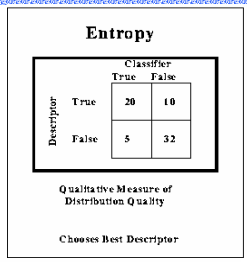
Como Trabalham as Árvores de Decisão

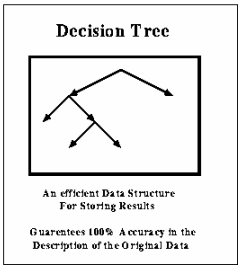
Vamos exemplificar para o caso do algoritmo ID3.

Principais conceitos:

- **entropia - utilizada para encontrar o parâmetro mais significativo na caracterização da classificação;**
- **árvore de decisão - modo eficiente, interpretável e intuitivo de organizar os descritores que podem ser utilizados como função preditiva.**

O ID3 encontra os preditores e os seus valores de divisão na base do ganho em informação que essas divisões proporcionam.





Como Trabalham as Árvores de Decisão

Ganho significa aqui a diferença entre a quantidade de informação necessária para efectuar uma predição correcta antes e depois da divisão ser feita.

- **se a quantidade de informação for muito menor, depois da divisão, então terá diminuído a desordem relativamente ao único segmento original;**
- **ganho é definido, como a diferença entre a entropia do segmento original e as entropias acumuladas dos segmentos resultantes da divisão.**

A entropia dos segmentos folhas são acumuladas pesando a sua contribuição para a entropia da divisão efectuada, de acordo com o número de registos que contem.

Como Trabalham as Árvores de Decisão

	Divisão Esquerda	Divisão Direita
Divisão A	+ + + + -	+ - - - -
Divisão B	+ + + + + - - - -	-

Mas este segmento de entropia = 0 tem apenas um registo, e dividir um registo de cada vez, não criará uma árvore de decisão muito útil.

Um segmento com 1 registo não deverá proporcionar padrões úteis repetíveis.

A divisão A é muito melhor do que a B, porque separa mais os dados, apesar da divisão B criar um novo segmento que é perfeitamente homogéneo (entropia 0)

Assim, o cálculo (métrica) a utilizar para determinar a divisão escolhida, deve levar em conta dois efeitos principais:

- **quanto baixou a desordem nos novos segmentos?**
- **como seria pesada a desordem em cada segmento?**

Métrica Utilizada em ID3

Métrica de Entropia (derivada dos trabalhos de Claude Shannon e Warren Weaven, sobre teoria de informação)

A equação é: $-\sum p \lg_2 (p)$

p - probabilidade do valor de predição ocorrer num nó particular da árvore

(terá um valor mínimo de 0 e máximo de 1, e assim a entropia terá um valor mínimo de 0 e máximo de 1)

Para verificar:

- **se tivermos um segmento com 100 clientes, em que 100 não renovam e 0 renovam, teremos uma entropia 0 (mínimo de desordem):**
- **$-1.0 * \lg_2(1.0) + -0.0 * \lg_2(0.0) = 1 * 0 + -0.0* -inf. = 0$**
- **para os mesmos 100 clientes, se tivermos 50 que renovem e 50 que não renovem, teremos um máximo de desordem:**
- **$-0.5*\lg_2(0.5) + - 0.5* \lg_2(0.5) = -0.5*-1 + - 0.5*-1 = 1$**

Métrica Utilizada em ID3

	Divisão Esquerda	Divisão Direita	Entropia Esquerda	Entropia Direita
Divisão A	+ + + + -	+ - - - -	-4/5lg(4/5) + -1/5lg(1/5)=0.72	-1/5lg(1/5) + -4/5lg(4/5) =0.72
Divisão B	+ + + + + - - - -	-	-5/9lg(5/9) + -4/9lg(4/9)=0.99	-1/1lg(1/1) + -0/1lg(0/1)=0

Entropia total:
Divisão A = 0.72+0.72 = 1.44

Entropia total:
Divisão B = 0.99+0.0 = 0.99

Conclusão:

Com esta medida de entropia, a divisão B seria muito melhor do que a A, contrariando o dito atrás (o facto de na divisão B, termos um segmento de um só registo, o que se fosse geral, não levaria a divisões úteis)

Métrica Utilizada em ID3

	Divisão Esquerda	Divisão Direita	Entropia Esquerda	Entropia Direita
Divisão A	+ + + + -	+ - - - -	-4/5lg(4/5) + -1/5lg(1/5)=0.72	-1/5lg(1/5) + -4/5lg(4/5) =0.72
Divisão B	+ + + + + - - - -	-	-5/9lg(5/9) + -4/9lg(4/9)=0.99	-1/1lg(1/1) + -0/1lg(0/1)=0

Peso: 5 registos em 10

Pesar a desordem em cada segmento, com o número de registos em cada segmento resultante da divisão:

Divisão A = 1/2 * 0.72 + 1/2 * 0.72 = 0.72

Divisão B = 9/10 * 0.99 + 1/10 * 0 = 0.89

Conclusão:

já seria escolhida a divisão A

É ainda introduzida uma nova melhoria na métrica, que leva em conta a cardinalidade do predictor - Rácio do Ganho - que será menor se a cardinalidade do predictor for alta (para evitar a formação de segmentos pequenos em predictores de alta cardinalidade)

Cálculo da Entropia em ID3 (1)

Consiste em três fases:

1 - Entropia antes do teste - a entropia é calculada para o data set (relativo ao nó em análise);

2 - Entropia de cada ramo - a entropia é calculada para cada um dos subconjuntos formados pela divisão do set principal, com respeito a descritor;
Somatório das Entropias dos ramos - as entropias calculadas são multiplicadas pelo tamanho do subconjunto com respeito ao data set inicial. São então adicionadas para formar a entropia resultado da divisão;

3 - Ganho de entropia - a diferença entre a entropia total original e a entropia dos ramos. Esta é a medida do quão significativo o descritor é relativamente ao classificador.

Cálculo da Entropia em ID3 (2)

Tabela de Distribuição
(para descritor binário)
C é um Classificador binário (valores= true, false)
D é um Descritor binário (valores true, false)

	D verdade	D falso	Totais
C verdade	tt	tf	ct
C falso	ft	ff	cf
Totais	dt	df	t

Cálculo de Entropia para divisão relativamente ao descritor D

1 - Entropia antes :

$$E = -(ct/t)\ lg(ct/t) - (cf/t)\ lg\ (cf/t)$$

2 - Entropia dos Ramos

Entropia de (D é verdade):

$$e1 = -(tt/dt)\ lg(tt/dt) - (ft/dt)\ lg(ft/dt)$$

Entropia de (D é falso):

$$e2 = -(tf/dt)\ lg(tf/dt) - (ff/dt)\ lg(ff/dt)$$

Entropia da divisão:

$$e = (dt/t)\ e1 + (df/t)\ e2$$

3 - Ganho de entropia

$$G = E - e$$

C4.5 Melhora ID3

Melhora o desempenho do ID3 nas seguintes áreas:

- **podem ser utilizados predictores com valores em falta;**
- **podem ser utilizados predictores com valores contínuos;**
- **é introduzida a poda;**
- **podem ser derivadas regras.**

CART

Algoritmo de exploração e predição desenvolvido mas Univ. de Stanford e Berkeley.

- **a selecção de cada predictor é baseada na sua habilidade de separar os registos de diferentes predições**
- **uma das medidas a utilizar é a métrica de entropia já vista no ID3**
- **outras métricas:**
  - **índice de diversidade Gini** $1 - \Sigma$ **(probabilidade para cada predição)^2, pesada pela probabilidade do segmento**
    - **Será maior quando a proporção de cada valor da predição forem equivalentes (maior entropia) e menor (zero) quando o segmento for homogéneo.**
  - **métrica twoing, similar ao Gini, mas tende a favorecer divisões mais balanceadas (segmentos com tamanhos mais equivalentes) - evitando o problema dos nós "dangling" – a divisão escolhida cria um segmento muito pequeno, enquanto que a maioria dos registos fica num segmento filho quase igual ao original**

CART

- **faz uma divisão em predictores não ordenados (ex. cor dos olhos) impondo uma ordem aos valores;**
- **efectua a validação automática da árvore:**
  - **inicia-se pela criação de uma árvore muito complexa**
  - **segue-se a fase da poda, criando uma árvore geral óptima, com maior probabilidade de trabalhar bem em dados novos.**
- **suporta o manuseamento de dados em falta,**
  - **quando a árvore está a ser criada, não considerando o registo particular na determinação da divisão óptima**
  - **na fase de predição, valores em falta, são tratados via substitutos (ex. tamanho do sapato para substituir altura)**

CHAID

CHAID - Chi Square Automatic Interaction Detector

- **para selecção das divisões, em vez de utilizar a entropia ou métrica Gini, utiliza a técnica do Qui-quadrado, utilizando tabelas de contingência para determinar que predictor categórico está mais longe da independência face aos valores de predição;**
- **todos os predictores devem ser categóricos ou transformados em categóricos via "binning", dado que se baseia em tabelas de contingência;**
- **embora o binning possa ter consequência perniciosas, o desempenho actual do CART E CHAID são comparáveis em modelos reais utilizados em direct marketing.**

Tendências Futuras em Árvores de Decisão

Continua a investigação para melhorar os algoritmos de árvores de decisão:

algoritmo C5.0

- **inclui "boosting" - técnica que combina árvores de decisão múltiplas num classificador simples.**

um outro produto - MineSet

- **permite que árvores de decisão múltiplas (ou subárvores) coexistam como opções. Cada opção faz uma predição e depois vota-se a predição consensual.Estas técnicas tratam a questão dos problemas de suboptimização resultantes do aspecto "greedy" dos algoritmos de árvores de decisão.**

Outros trabalhos incluem suporte de variáveis contínuas e árvores oblíquas: árvores com divisões multivariáveis.

Regras de Associação

Regras de Associação

Trata-se da forma de Data Mining que de mais perto se assemelha ao processo que a maioria das pessoas lhe associa "Minar uma grande base de dados à procura da pepita de ouro"

Aqui a pepita de ouro significa:

"Uma regra que diga algo sobre a base de dados que não se saiba e que provavelmente não sejamos capazes de articular explicitamente, além de ser algo de interesse"

Regras de Associação

Os gestores de marketing gostam de regras como:

"90% das mulheres possuidoras de carros de desporto vermelhos e cães pequenos, usam Chanel Nº5"

- **Este tipo de regras descritivas mostra perfis claros dos clientes que poderão constituir alvos das suas acções de marketing.**

Estas regras são chamadas de Regras de Associação, sendo um dos outputs possíveis de várias técnicas de Data Mining.

- **as regras de associação são sempre definidas em termos de atributos binários**
- **é possível encontrar associações em grandes bases de dados**

mas

- **além de associações de interesse**
- **encontrar-se-ão muitas associações de pouco valor, já que o número de associações será quase infinito.**

Regras de Associação

Sob a forma de regras relativamente simples, como:

⇨ **Se forem adquiridas baguetes, então creme de queijo sê-lo-á também 90% das vezes e esse padrão ocorrerá em 3 % de todos os cestos de compras.**

⇨ **Se forem adquiridas plantas então será adquirido adubo 60% das vezes e esses dois itens serão adquiridos em conjunto em 6% dos cestos de compras.**

Força - Retira todas as possíveis regras de interesse (não deixa nenhuma minério por peneirar)

Fraqueza - O utilizador pode facilmente ver-se submergido com tão grande quantidade de regras, sendo difícil avaliá-las todas (seria necessária uma 2.ª passagem DM para encontrar as mais valiosas pepitas de entre as encontradas).

Regras de Associação - Valor p/ o Negócio

Medida de DM	Descrição
Automatização	Tendem a ser altamente automatizadas na construção, ordenação e apresentação das regras. Muitas vezes é pedido ao utilizador a avaliação de cada uma de muitas regras por forma a determinar a importância de cada uma.
Clareza	As regras são geralmente simples e fáceis de compreender, embora o porquê da sua ocorrência possa não ser fácil de saber. Dado o grande número de regras, alguma clareza pode perder-se e o utilizador pode ver-se submergido com regras obscuras que não façam sentido ou regras óbvias, já conhecidas.
ROI	Embora possam ser utilizadas para predição, são quese sempre utilizadas para aprendizagem não supervisionada para encontra algo não conhecido. É mais difícil de quantificar o ROI das regras "interessantes" do que avaliar do ROI quando temos um problemas de predição com alvo bem conhecido.

Regras de Associação

Problema:

- **Separar a informação válida do mero ruído**

Solução:

- **Introduzir algumas medidas para distinguir associações com possível interesse daquelas sem interesse**

Uma regra de associação representa-se:
Revista_Música, Revista_Casas --> Revistas_Carros

Descrição: alguém que lê revistas de música e de casas, será, possivelmente, um leitor de revistas de carros.

O que é uma Regra?

"Se são comprados pickles, então também é comprado Ketchup"

Antecedente: pode ser uma ou mais condições, que devem ser todas verdadeiras por forma ao consequente seja verdadeira, dada a precisão

Consequente: Geralmente só uma condição e não uma combinação múltipla.

Para uma regra ser útil, além da própria regra, há que fornecer dois suplementos:

1. **Precisão - Quantas vezes está correcta a regra? É a probabilidade de que se o antecedente for verdadeiro, então o consequente será verdade. Precisão alta, significa estar em presença de uma regra de alta confiança, daí ser também chamada de confiança.**
2. **Cobertura - Quantas vezes se aplica a regra? Relacionado com a % da base de dados que é coberta pela regra. Alta cobertura significa que a regra se aplica muitas vezes e que também é pouco provável que se trate de algo espúrio introduzido pela técnica de amostragem ou ideosincrasias.**

Exemplos de Precisão e Cobertura de Regras

Regra	Precisão (%)	Cobertura (%)
Se forem comprados cereais p/ pequeno almoço, então será comprado leite	85	20
Se for comprado pão, então será comprado queijo suíço	15	6
Se o cliente tiver 42 anos e comprar salgadinhos e amendoins torrados, então cerveja será comprada	95	0.01

A precisão e a cobertura são ambos importantes na determinação da utilidade duma regra. A primeira regra é um padrão que ocorre muito e é quase sempre verdade. A a terceira regra quase nunca está errada, mas também quase não é aplicável.

Como Avaliar uma Regra

Em Estatística:

<u>Cobertura</u> - É simplesmente a probabilidade *a priori* da ocorrência do antecedente

<u>Precisão</u> - É a probabilidade condicional do consequente no antecedente

Ex. Num supermercado a regra "compra de leite implica a aquisição de ovos"

T = 100 = n.º total de cestos de compras na base de dados

E = 20 = n.º de cestos de compras com ovos

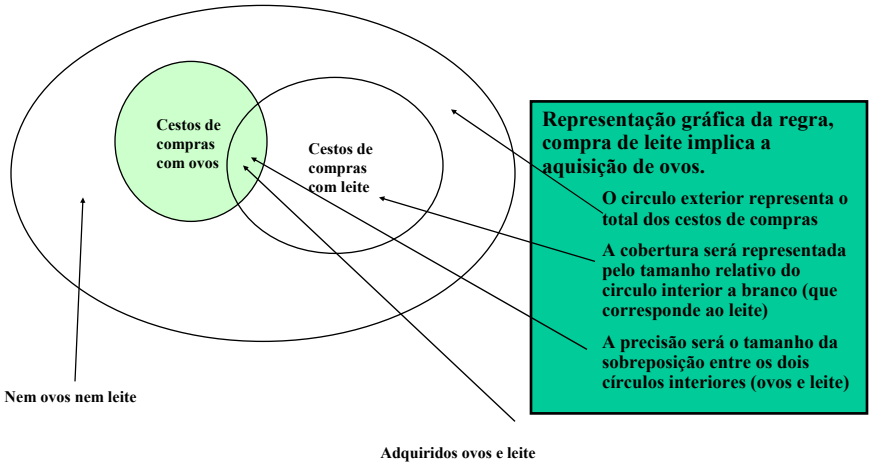
M = 40 = n.º de cestos de compras com leite

B = 20 = n.º de cestos de compras com ovos e leite

<u>Precisão</u> = n. de cestos de compras com ovos e leite / n.º de cestos de compras com leite = 20 / 40 = 50%

<u>Cobertura</u> = n.º de cestos de compra com leite / n.º total de cestos de compra = 40 / 100 = 40%

Regras (Representação Gráfica)



O que Fazer com uma Regra?

1. <u>Alvo o Antecedente</u> - regras que tenham um dado valor no antecedente são procuradas e mostradas ao utilizador. Ex. um armazém de ferramentas pode pedir todas as regras que contenham pregos, parafusos e porcas no antecedente por forma a tentar saber as consequências que poderão advir na aquisição de produtos de margem elevada, se os primeiros forem descontinuados - caso: pessoas que adquirem pregos, adquirem também martelos caros.
2. <u>Alvo o Consequente</u> - Conhecer o que vai afectar o consequente - Pode ser útil conhecer regras que contenham café no consequente e assim colocar o café e o item assim encontrado perto, aumentando as vendas de ambos.
3. <u>Alvo baseado na Precisão</u> - Algumas vezes a coisa mais importante para o utilizador é a precisão da regra. Regras com grande precisão (90-95%) implicam relacionamentos fortes que podem ser explorados, mesmo se tiverem cobertura baixa.
   Ex. uma regra que tenha 0.1% de cobertura, mas precisão de 95%, poderá ser aplicada uma em cada 1000x, mas terá grande probabilidade de se verificar, entaõ se der muito lucro, poderá ser uma regra valiosa.

O que Fazer com uma Regra?

4. <u>Alvo baseado na Cobertura</u> - Olhar para as regras ordenadas por cobertura. Pode assim ficar-se com uma visão de alto nível do que acontece na base de dados na maior parte do tempo. Trata-se do conjunto de regras mais prontas aplicar.
5. <u>Alvo baseado no Interesse</u> - As regras serão interessantes quando tiverem cobertura e precisão alta e se desviarem da norma. Tem havido muitas formas de ordenar essas regras por alguma medida de interesse, e assim poder pesar-se entre a cobertura e precisão.

<u>Interesse</u> - Trata-se de uma medida bastante difícil de estimar, pois que as regras mostram relacionamentos entre os valores nos predictores da base de dados e não um alvo bem definido para predição.

Descoberta com Regras

Proporcionam:

1. **Uma visão detalhada dos dados onde ocorrem padrões significativos um pequeno número de vezes - Micro Nível - regras fortes que cobrem poucas situações, mas podem aplicar-se a algo muito valioso (clientes lucrativos, p.ex.), ou representam um pequeno subgrupo, mas em crescimento, indicando uma deslocação de mercado.**
2. **Uma visão genérica dos dados e padrões globais contidos na base de dados - Macro-Nível - Padrões que cobrem muitas situações podem ser utilizados muitas vezes e com grande confiança; também podem ser utilizados para sumarizar a base de dados.**

Predição com Regras

Uma regra pode efectuar predição - o consequente é o alvo e a precisão da regra será a precisão da predição.

Mas...

como há muitas regras, pode haver conflitos e inconsistências dadas as precisões diversas.

- **Como calcular a precisão combinada?**

Antecedente	Consequente	Precisão (%)	Cobertura (%)
Baguetes	Creme de Queijo	80	5
Baguetes	Sumo de Laranja	40	3
Baguetes	Café	40	2
Baguetes	Ovos	25	2
Pão	Leite	35	30
Manteiga	Leite	65	20
Ovos	Leite	35	15
Queijo	Leite	40	8

Precisão *versus* Cobertura

	Precisão Baixa	Precisão Alta
Alta Cobertura	A regra raramente está correcta, mas pode ser utilizada muitas vezes	A regra está correcta muitas vezes e pode ser utilizada amiúde
Baixa Cobertura	A regra raramente está correcta e só pode ser utilizada muito raramente	A regra está correcta muitas vezes, mas só pode ser utilizada raramente

Definição de Interesse

O interesse deverá ter a ver com a precisão e a cobertura, mas deverá possuir mais algumas características:

1. O interesse será = 0 se a precisão for igual à precisão geral (probabilidade *a priori* do consequente)

Antecedente	Consequente	Precisão (%)	Cobertura (%)
Sem constrangimentos	Então o cliente sairá	10	100
Se saldo cliente > 3000€	Então o cliente sairá	10	60
Se o cliente tem olhos azuis	Então o cliente sairá	10	30
Se o n. Seg. social = 151 555 345	Então o cliente sairá	100	0.000001

2. O interesse aumenta à medida que a precisão aumenta e inversamente, mantendo a cobertura
3. O interesse aumenta aumenta ou diminui com a cobertura para uma dada precisão
4. O interesse diminui com a cobertura para um certo número de respostas correctas (dado que a precisão = n.º de respostas correctas / cobertura)

Outras Medidas úteis em Regras

- **Suporte**: quão frequentemente as regras (antecedente e consequente) ocorrem;
  ex. o suporte será a percentagem de registos da base de dados de subscritores de revistas que verifiquem a regra
  revista_música, revista_casa --> revista_carros
  ou seja, a percentagem de subscritores que lêem as três revistas.
- **Significância**: diz-nos quão diferente o padrão mostrado na regra é comparado com a ocorrência aleatória de fenómenos independentes;
- **Simplicidade**: ajuda ao utilizador na construção de intuições e confirmação da regra via intuições;
- **Novidade**: ajuda ao utilizador a encontrar regras que ocupam regiões no espaço de predição onde estão novas regras.
- **Confiança**: mostra o grau de força da associação, ou seja, mede a ligação entre os registos à esquerda e à direita da regra (semelhante à precisão).
  Ex. no caso da regra revista_música, revista_casa--> revista carros
  será a percentagem dos registos de leitores de revistas de música e casas que tb. lêem revistas de carros.

Como funciona a Indução de Regras

Embora os algoritmos variem, têm muitos passos comuns:

1. Pré-processamento dos dados por forma a que cada predictor tenha intervalos bem-definidos em vez de valores contínuos (estes serão os constrangimentos que serão adicionados às regras um de cada vez para serem criadas regras mais complexas);
2. Gerar regras iniciais para dados com um constrangimento apenas;
3. A partir dos registos, gerar regras que tenham um constrangimento adicional à regras dadas;
4. Manter o grupo de regras que sejam boas candidatas de terem constrangimentos adicionais;

Como funciona a Indução de Regras

5. Continuar a adicionar constrangimentos às regras até que o critério de paragem tenha sido atingido para todas as regras (normalmente algum patamar mínimo de precisão, cobertura ou suporte);
   - adicionar constrangimentos pode no máximo não afectar a cobertura e suporte, mas o mais provável é que diminuirão decerto os seus valores, nunca decerto os aumentarão; se uma regra já tiver um mínimo de cobertura ou suporte, pode não fazer qualquer sentido expandi-la, pois que, mesmo se a precisão for alta, há um valor abaixo do qual a regra não revelará qualquer valor - se a regra nunca é utilizada, não interessa que precisão poderá ter;
6. Organizar as regras com base na sua utilidade (precisão, suporte, confiança, significância, simplicidade e novidade).

Como funciona a Indução de Regras

Em resumo:

- A geração começa simplesmente por tomar cada valor de cada predictor de cada registo e emparelhá-lo com todos os outros valores dos predictores
- Estes pares representam a primeira passagem de regras simples if-then
- Destas é possível determinar que regras são boas candidatas para expansão para novos constrangimentos (passar algum patamar mínimo de precisão e cobertura)
- Continuar o processo até que nenhuma regra passe o patamar

Algoritmo Força-Bruta p/ Geração de Regras

1. Gerar todos os pares de predictores/valor para cada registo como 1.º conjunto de regras
2. Contar o n.º de ocorrências para cada regra e o antecedente por si
3. Calcular a precisão, cobertura e suporte e eliminar as regras que não satisfaçam o limiar mínimo requerido
4. Para cada registo, ver que regras se lhe aplicam e adicionar a cada regra um constrangimento adicional (valor de predictor) do registo
5. Voltar ao passo 1 até que as regras se tornem demasiado complexas ou que o patamar p/ a cobertura mínima ou suporte não seja satisfeito

Algoritmo Força-Bruta p/ Geração de Regras

- O algoritmo anterior, é computacionalmente caro, mas pode ser eficientemente implementado via algoritmos de sort ou hashing, muito bem adaptados a processamento paralelo
- Podem ser utilizados métodos heurísticos para saber que constrangimentos novos serão os elegíveis para cada regra, mas o crescimento do desempenho dos computadores, torna o algoritmo realizável

Regras de Associação - Forças

- produz regras fáceis de perceber
- relativamente fácil de desenvolver
- muito úteis em descoberta e predição
- não sensível a valores em falta ou ruído, pois que são extraídos todos os padrões de interesse
  - semelhantes à tomada de decisão por comités - muitos membros votam e a maioria oferece a decisão
  - todas as regras que contribuem para a predição final asseguram que uma só regra ou valor em falta ou passo em falso de um algoritmo greedy, não levarão a incorrecção na predição

Regras de Associação - Limitações

A procura e descoberta de associações é muito lata:

- actualmente não há algoritmo algum de data mining que nos dê, de forma automática, tudo o que seja de interesse relativo a uma dada base de dados.
- temos de ter já uma ideia geral do que procuramos, pois que:
  - um algoritmo que encontre muitas regras úteis, encontrará, decerto, montanhas de regras inúteis;
  - já um algoritmo que encontre um número limitado de associações, sem ajuda de sintonia fina,decerto passará em falso, informação interessante.
- Dada a decisão por consenso, assiste-se ao obscurecimento da simplicidade da regra individual que mostra a probabilidade condicional da decisão ser tomada

Construção de Regras (sobreadaptação)

Quando o suporte a uma regra fica demasiado pequeno

- **há poucos registos na B.D. que consubstanciem a regra**
- **a própria regra pode ser inteiramente espúria (resultado de ruído ou variações estatísticas devidas à pequenez da amostra)**

Se o algoritmo, no limite, não parar por falta de suporte

- **o sistema criará regras que contêm só um valor para cada predictor**
- **um só registo a ser coberto por cada regra**
- **fenómeno de sobreadaptação**
- **produção de regras não gerais**

Na realidade o limiar requerido para o suporte e cobertura:

- **são tentativas de limitar a sobreadaptação**
- **leva a produzir regras mais gerais, utilizáveis em novas situações**

Regras *versus* Árvores de Decisão

As árvores de <u>decisão</u>, como vimos, produzem <u>regras que são mutualmente exclusivas e colectivamente exaustivas</u> com respeito à base de dados de treino , enquanto que os <u>sistemas de regras de indução</u> produzem <u>regras não mutualmente exclusivas e que podem ser colectivamente exaustivas</u>

- **os algoritmos de indução de regras vão de baixo para cima e coligem todos os padrões possíveis que possam ser interessantes, utilizando-os depois para alguns alvos de predição**
- **as árvores de decisão trabalham com um alvo de predição à vista, de cima para baixo, numa procura "greedy", procurando a melhor divisão no próximo passo (não olhando mais do que o próximo passo)**

Ofertas Correntes e Melhoramentos Futuros

Disponíveis:

Information Discovery
Attar Software
IBM e outros

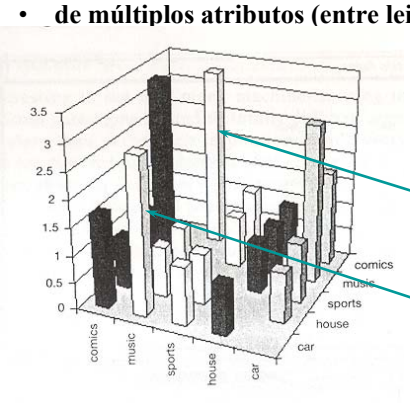
No futuro:

- **melhor visualização das regras**
- **cluster de regras e produção de hierarquias**
- **avanços em:**
  - **saber das regras efectivamente interessantes (difícil, pois que é tb. uma arte)**
  - **algoritmos mais deterministas e embebidos no próprio DW**
- **surgir de noção mais genérica dos aspectos comuns a algoritmos de predição e descoberta e assim desvanecer a separação entre técnicas de NN, regras e árvores de decisão.**

Regras de Associação - Tipos de Associações

As associações podem ser:

- **binárias - entre cada 2 tipos de leitores de revistas (um exemplo apresentado abaixo)**
- **de múltiplos atributos (entre leitores de vários tipos de revistas)**

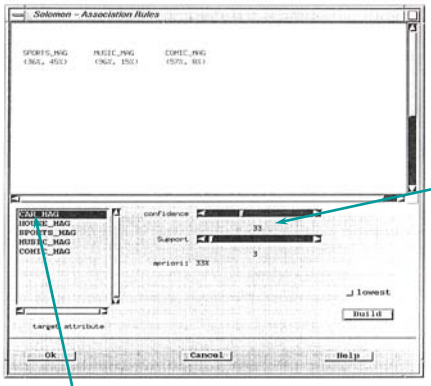


Tem 25 possíveis regras de associação p/ associações binárias

Para associações entre múltiplos atributos, o número crescerá exponencialmente

Vemos que há uma correlação alta entre leitores de revistas de humor e música e entre leitores de revistas de carros e música

Regras de Associação - Exemplo



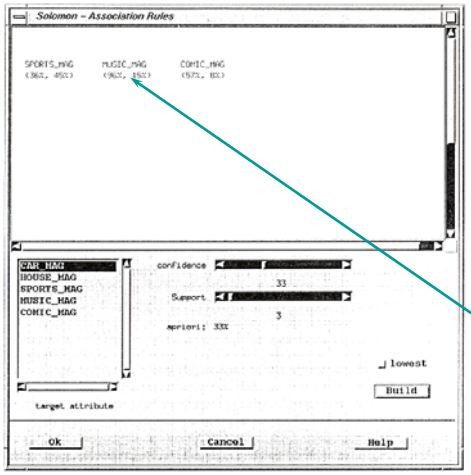
Os níveis de suporte e confiança foram colocados a 3% e 33%, respectivamente, ou seja, não estamos interessados em subgrupos < 3% da base de dados e, de entre esses, só queremos associações que sejam verificadas em pelo menos 33% dos registos.

Como se vê, a ferramenta proporciona um ambiente interactivo que permite conhecer detalhes de conjuntos de regras consideradas interessantes.

Neste caso, seleccionámos a revista_carros como atributo alvo: isto é, estamos interessados em leitores de revistas de carros, vamos investigar possíveis associações desses leitores a outras revistas

Regras de Associação - Exemplo

1ª fase da nossa investigação: todos os atributos relevantes são investigados



para a revista de casas não é encontrada associação, dados os níveis de confiança e suporte especificados

são encontrados 3 grupos de associações:

revista_desporto--> revista_carros

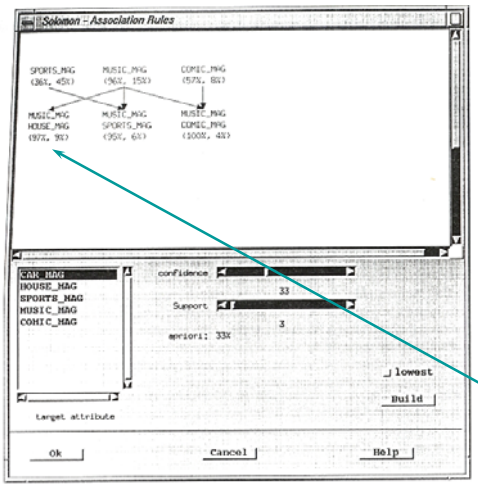
revista_música--> revista_carros

revista_humor--> revista_carros

a mais interessante, pois que tem um nível de confiança de 96% com suporte de 15%

Regras de Associação - Exemplo

2ª fase: investigar as regras que parecem mais interessantes, neste caso, a segunda



encontram-se 3 novas regras:

revista_música, revista_casas --> revista_carros

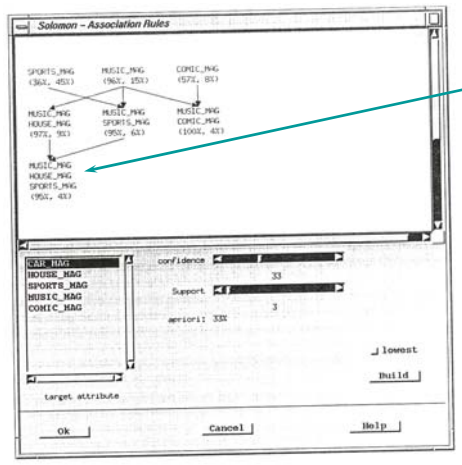
revista_música, revista_desporto --> revista_carros

revista_música, revista_humor --> revista_carros

a mais interessante, pois que tem um nível de confiança alto (97%) com um suporte ainda alto (9%)

Regras de Associação - Exemplo

3ª fase: expandir novamente a regra interessante atrás encontrada



obtem-se:

revista_música, revista_casas, revista_desporto --> revista_carros

mas o nível de confiança e suporte não aumenta, não se consegue melhor do que o caso anterior

Parece que o melhor que se pode obter é: todos os leitores de revistas de música e casas, são também leitores de revistas de carros com uma confiança de 97% e suporte de 9%.