



UNIVERSIDADE DOS AÇORES
-DEPARTAMENTO DE MATEMÁTICA-
MONOGRAFIA

Árvores de Classificação

Aluno:

Marco António dos Santos Rodrigues

2004/2005

Índice

<i>Introdução</i>	<i>1</i>
<i>Capítulo I – Métodos de Classificação</i>	<i>2</i>
1. <i>Aprendizagem Supervisionada e não Supervisionada</i>	<i>3</i>
2. <i>Teoria de Decisão de Bayes</i>	<i>5</i>
3. <i>Projection Pursuit</i>	<i>6</i>
4. <i>Análise de Clusters</i>	<i>7</i>
6. <i>Modelos de Mistura</i>	<i>9</i>
<i>Capítulo II – Aspectos Genéricos das Árvores de Classificação</i>	<i>12</i>
1. <i>Árvores de Decisão / Classificação: Fundamentos</i>	<i>12</i>
2. <i>Crescer Árvores de Decisão / Classificação</i>	<i>14</i>
3. <i>Um exemplo de uma Árvore de Classificação</i>	<i>15</i>
<i>Capítulo III – Principais Algoritmos de Árvores de Classificação e Decisão</i>	<i>18</i>
1. <i>Algoritmos ID3 e C4.5</i>	<i>19</i>
2. <i>Algoritmo CART</i>	<i>22</i>
4. <i>Algoritmo CHAID</i>	<i>24</i>
5. <i>Algoritmo QUEST</i>	<i>25</i>
<i>Capítulo IV – Caso de Estudo</i>	<i>26</i>
<i>Conclusão</i>	<i>30</i>
<i>Bibliografia</i>	<i>31</i>

Introdução

Actualmente, o ambiente que rodeia as decisões de carácter financeiro ou de gestão tendem a ser cada vez mais exigentes. Com o desenvolvimento das técnicas quantitativas, cresceram igualmente os dados armazenados, justificando deste modo a importância crucial do desenvolvimento de tecnologias de análise de dados com o objectivo de obtenção de informação e conhecimento sobre o mercado e a concorrência. As tecnologias de informação adoptadas pelas várias empresas apresentam-se como factor preponderante e determinante de sucesso. A análise de dados poderá ter como objectivos descobrir o que é realmente relevante para uma tomada de decisão.

Uma importante técnica utilizada em análise de dados é precisamente as Árvores de Decisão / Classificação usadas extensivamente pelas tecnologias de *Data Mining*. O conceito de *Data Mining* apresenta-se integrado num conjunto de metodologias de aprendizagem que se caracteriza, em particular, pela adaptação às grandes dimensões das bases de dados sobre as quais se apreende e extrai conhecimento.

As Árvores de Decisão / Classificação são representações simples do conhecimento e um meio eficiente de construir classificadores que estabelecem classes baseadas nos atributos de um conjunto de dados. Esta técnica é utilizada com êxito em diversos campos, tais como: no diagnóstico médico; na análise de mercados, na agricultura entre muitas outras.

Este trabalho está estruturado em Introdução, 4 Capítulos e Conclusão. O primeiro capítulo descreve sucintamente vários métodos de classificação; o segundo capítulo aborda as Árvores de Decisão ou Classificação de forma prática, constituindo a base deste trabalho; o terceiro salienta os vários algoritmos utilizados nas Árvores de Decisão/Classificação e o último capítulo apresenta um caso de estudo onde são postos em prática os conhecimentos adquiridos ao longo da realização do trabalho teórico.

Capítulo I – Métodos de Classificação

A tarefa de agrupar ou classificar objectos, em categorias, é uma das actividades mais comuns e primitivas do Homem e vem sendo intensificada, em função do grande volume de informações disponíveis actualmente, sobre as mais diversas áreas.

Para realizar essa tarefa, existem inúmeros métodos. Serão abordados neste trabalho de forma resumida os seguintes métodos: Teoria de decisão de Bayes; *Projection Pursuit*; Modelos de Mistura e Análise de *Clusters*, e de forma mais aprofundada: As Árvores de Decisão / Classificação, que determinam a base deste trabalho.

Para uma melhor compreensão sobre os vários algoritmos utilizados em classificação, introduz-se abaixo uma árvore que esquematiza as relações de funcionamento e utilização de vários métodos e modelos.

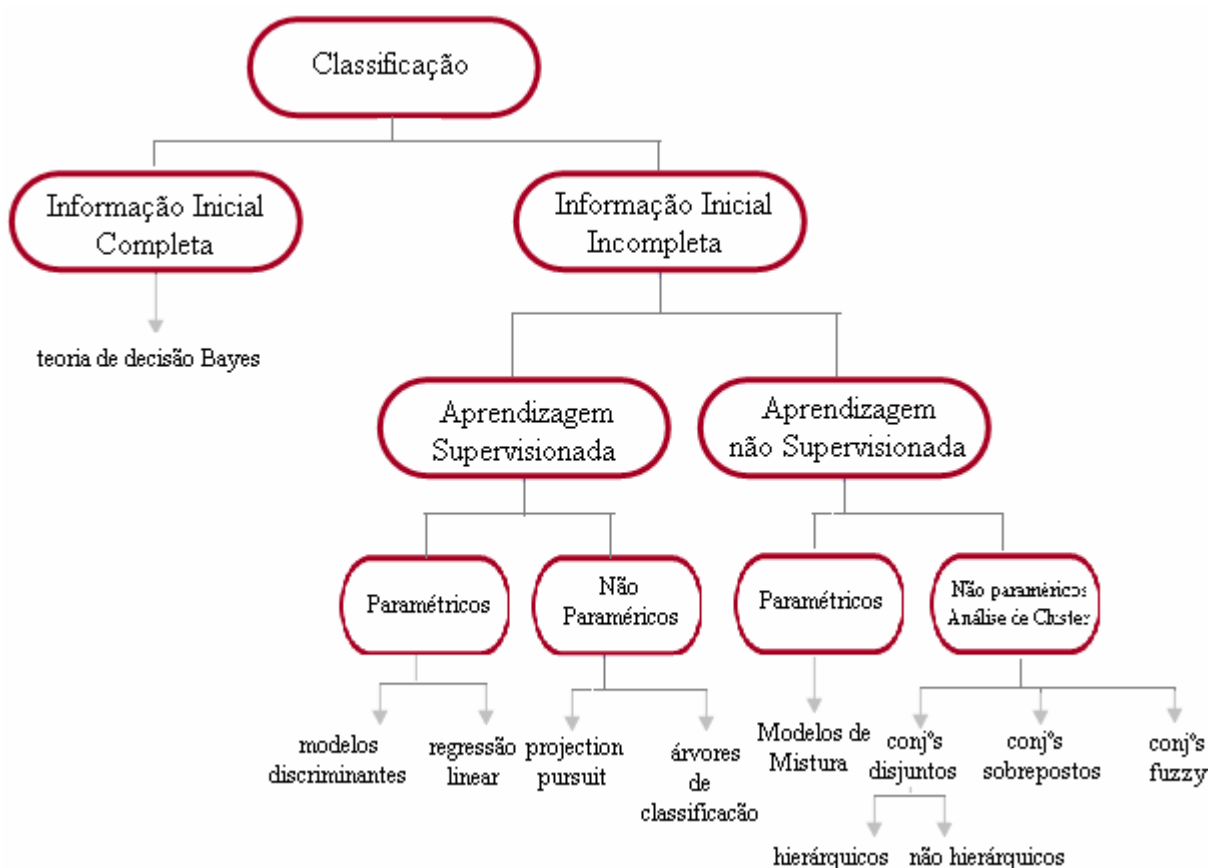


Figura 1 – Esquema representando diferentes métodos e modelos de classificação.

Apesar de existirem diversas denominações sobre o conceito de “classificação” podemos defini-la como sendo um processo de discriminação de unidades concretas ou abstractas em classes ou categorias, ou de forma abreviada, acto, efeito ou processo de distribuir por classes. Se estas classes estiverem definidas à partida e existir informação sobre a probabilidade de um determinado objecto pertencer a uma classe, entende-se como informação inicial completa, caso contrário, esta informação inicial estará incompleta. De acordo com o tipo de informação disponível o passo seguinte passa pela utilização e escolha de um método. Os métodos podem ser classificados em paramétricos e não paramétricos.

1. **Métodos paramétricos:** nos métodos paramétricos a distribuição da população ou do processo subjacente às observações tem uma dada forma e as inferências, condicionadas por esse pressuposto, dizem respeito a um ou a vários (em número finito) parâmetros (são exemplos a Regressão Linear, os Modelos Discriminantes e os Modelos de Mistura).
2. **Métodos não-paramétricos:** em geral, a forma da distribuição da população não é conhecida e as inferências processam-se em quadro muito menos restrito e muitas vezes não envolvem parâmetros (por exemplo as Árvores de Decisão, as técnicas de *Projection Pursuit* e *Análise de Cluster*).

1. Aprendizagem Supervisionada e não Supervisionada

A distinção principal em relação ao paradigma de aprendizagem, que é válido para todo o tipo de sistemas com capacidade de adaptação, é a aprendizagem supervisionada e aprendizagem não-supervisionada.

Na **aprendizagem supervisionada**, cada exemplo de treino está acompanhado por um valor alvo (*target*). Assim, utiliza-se uma variável dependente com informação sobre as classes a que pertencem cada uma das entidades da amostra de treino. Neste conceito, incluem-se técnicas da estatística multivariada como a regressão; análise discriminante; a regressão logística e novas técnicas da área de reconhecimento de padrões como as árvores de Decisão / Classificação e de regressão e as redes neuronais supervisionadas.

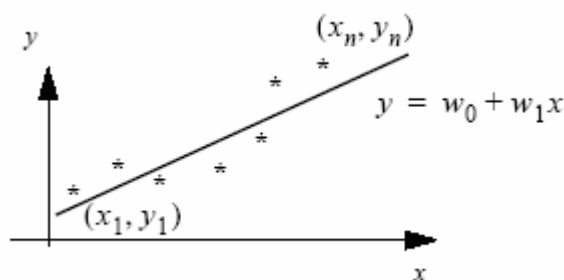


Figura 2 - Regressão Linear

Um exemplo de aprendizagem supervisionada é a regressão linear, utilizada para previsão. Usamos o caso unidimensional para facilitar a ilustração (Figura 1). Nesse problema, o conjunto de treino consiste em pares de números reais (x_p, y_p) . O objectivo da aprendizagem é a determinação de coeficientes w_0 e w_1 da recta $y = w_0 + w_1x$. O algoritmo de aprendizagem tenta minimizar a discrepância entre o valor da variável dependente y_p e o valor da resposta $\hat{y}_p = w_0 + w_1x_p$ do sistema.

O conjunto de treino para a aprendizagem supervisionada é utilizado de forma semelhante. Este conjunto de treino é, então, um conjunto de associações problema / solução. Assim sendo, na aprendizagem supervisionada o exemplo de treino já está classificado, correspondendo o passo de aprendizagem à estimação de modelos semelhantes ao que na estatística se denomina por análise discriminante.

O processo de **aprendizagem não supervisionada** é a aprendizagem por meio de observação e descoberta, ou seja, não existem à partida dados rotulados ou classificados. O número de categorias ou classes pode não estar definido *a priori*, neste caso o método tem que encontrar atributos estatísticos relevantes. Enquadram-se neste conceito: os algoritmos de análise de grupos (*clustering*), os modelos de mistura e as redes neuronais não supervisionadas.

Na área da medicina, pode constituir exemplo de aprendizagem não supervisionada a detecção de doenças a partir de imagens, por exemplo, imagens de raio-X. Existem várias áreas dentro da imagem que se devem atribuir ao mesmo material, como é o caso do osso. O número de materiais (isto é, de grupos) não é conhecido *a priori*. O objectivo do método é descobrir o número dos grupos diferentes e, ao mesmo tempo, associar cada ponto da imagem ao respectivo material (grupo). Os dados de entrada para o método seriam os pontos da

imagem e o valor da coloração ou nível de cinzento. A resposta ideal do método seria a disposição do grupo a que pertence essa região da imagem.

2. Teoria de Decisão de Bayes

A Teoria de Decisão de Bayes formula o problema de classificação num contexto probabilístico. O seu principal objectivo é determinar uma lei de decisão que minimize um critério de risco. A solução encontrada é óptima mas assume um conhecimento completo sobre a distribuição dos dados.

Consideremos a fórmula de Bayes, (REIS, *et al.* 1999), uma vez que traduz o princípio fundamental da aprendizagem *bayesiana*. Se $\{A_1, A_2, \dots, A_n\}$ é uma partição sobre Ω , então, para um novo conjunto de objectos B definido em Ω , com $P(B) > 0$ tem-se:

$$P(A_j | B) = \frac{P(A_j)P(B | A_j)}{\sum_{i=1}^n P(A_i)P(B | A_i)}, j = 1, 2, \dots, n. \quad (1)$$

designando $P(A)$ a probabilidade do acontecimento A e $P(A|B)$ a probabilidade de A condicionada por B , definida por:

$$\frac{P(A \cap B)}{P(B)}. \quad (2)$$

Uma consequência imediata desta fórmula é que para dois acontecimentos A e B , tais que $P(A) > 0$ e $P(B) > 0$, tem-se:

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}. \quad (3)$$

Supondo que o nosso conjunto de dados (instâncias de treino) é designado por D , então pela fórmula anterior temos forma de calcular a probabilidade de uma hipótese h pertencer a um dos grupos, tendo por base esses dados:

$$P(h | D) = \frac{P(D | h)P(h)}{P(D)} \quad (4)$$

em que a probabilidade $P(h/D)$ é a denominada “probabilidade *a posteriori*”, $P(h)$ a “probabilidade *a priori*” da hipótese h . Considerando um espaço de hipóteses possíveis H , o

objectivo consiste em determinar qual a melhor hipótese, tendo em conta os dados observados em D .

Se interpretarmos a melhor hipótese, como a mais provável, atendendo aos dados, isto é, a hipótese com melhor valor de “probabilidade *a posteriori*”, designada usualmente por h_{MAP} , (MAP – *Maximum a Posteriori Probability*) então o objectivo é:

$$h_{MAP} = \arg \max_{h \in H} P(h | D) \quad (5)$$

O classificador acima definido, traduz a maximização da distribuição *a posteriori* das classes, e é designado por classificador de máximo *a posteriori*.

Nos problemas de classificação, podemos aplicar a fórmula de Bayes como classificador, sendo o objectivo encontrar qual a classe mais provável, tendo em conta as instâncias de treino observadas. Se tivermos informação *a priori* completa tal procedimento é sempre possível de aplicar.

Desta forma compreende-se a importância que a abordagem *bayesiana* tem na aprendizagem indutiva, todavia, a sua aplicação prática é algo complicada, na medida em que é necessário conhecer as probabilidades exactas *a priori* habitualmente inacessíveis.

3. *Projection Pursuit*

A *projection pursuit* tem como principal objectivo encontrar projecções “interessantes” (tipicamente em uma ou duas dimensões) de um conjunto de dados multivariados, onde esse “interesse” é medido através de um índice de projecção.

Para melhor compreender como este método funciona, consideremos a Figura 2. Nesta figura projecta-se uma série de dados (tridimensionais) no espaço em duas dimensões. O processo é de simples compreensão - incidindo nos dados a tocha ligada indica a sombra nas telas (bidimensionais). Note-se como a sombra gerada pela tocha amarela desenha naturalmente dois grupos. Isto reflecte a separação no espaço tridimensional. A sombra gerada pela tocha vermelha indica somente um grupo. Para classificar em dois grupos seria preferível a sombra amarela à vermelha.

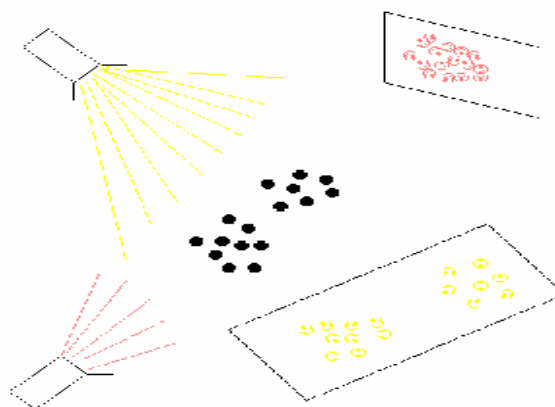


Figura 3 – Diversas projecções dos mesmos dados.

Se tivéssemos a tocha na posição vermelha quereríamos mover a tocha para a posição amarela para obter o retrato da sombra de dois grupos, por outras palavras, o objectivo seria perseguir o sentido da tocha amarela.

A designação de *projection pursuit* tem por base perseguir projecções adequadas em espaços com elevado número de dimensões e assim permitir a visualização de agrupamentos anteriormente ocultos.

4. Análise de *Clusters*

A análise de *clusters* constitui um processo pelo qual se procura classificar objectos em categorias não definidas à partida, como se verifica em todos os métodos não supervisionados.

Numa primeira fase, são necessárias medidas de similaridade ou dissemelhança, que serão usadas pelo algoritmo para unir num só grupo objectos com características similares. Essas medidas, em geral, podem ser classificadas em três tipos:

3. Medidas de Distância ou Dissemelhança;
4. Medidas de Semelhança;
5. Medidas de Paridade.

A expressão abaixo reproduz uma medida de distância, mais conhecida por distância Euclidiana entre dois pontos (i e j), isto é:

$$D_{ij} = \sqrt{\sum_{t=1}^p (X_{it} - X_{jt})^2} \quad (6)$$

em que X_i e X_j são as projecções dos pontos i e j na dimensão $t = \{1, 2, \dots, p\}$.

Como na maioria das vezes, as variáveis são medidas em diferentes unidades, a fórmula acima é frequentemente aplicada após a padronização para média zero e variância unitária (*z-score*).

Já nas medidas de paridade, os termos analisados são as semelhanças entre si, segundo um conjunto de atributos. Esses atributos apresentam escalas dicotómicas (binárias) e, após isso, essas características binárias são analisadas para construir métodos de paridade entre objectos.

A segunda fase, será a escolha do algoritmo a ser usado. Existem, de maneira geral, dois conjuntos de algoritmos: métodos hierárquicos e métodos não-hierárquicos, dentro dos grupos disjuntos.

Os procedimentos hierárquicos envolvem a construção de uma hierarquia em forma de árvore. Há, basicamente, dois tipos de métodos hierárquicos: aglomerativos e divisivos.

Nos algoritmos aglomerativos, cada objecto começa por corresponder a um cluster. Nos passos subsequentes, os dois *clusters* mais próximos são unidos num novo cluster. Esse processo é iterativo e realiza-se até que reste apenas um *cluster*.

O método hierárquico divisivo, inicia-se com um grande *cluster* agregando todos os objectos que, posteriormente, se vão dividindo e separando os objectos distantes do centro para formar outros *clusters* até o número de *clusters* desejados. No limite, o número de *clusters* formados é igual ao número de objectos a serem classificados, constituindo cada objecto o seu próprio cluster.

Os métodos não-hierárquicos, não envolvem um processo de construção do tipo árvore. Baseiam-se na obtenção de um número predefinido de *clusters* (k), que conterão todos os casos observados. Ou seja, em vez de se combinarem *clusters* semelhantes entre si, procura-se encontrar os k *clusters* que melhor solucionam o problema segundo a minimização ou maximização de uma medida de heterogeneidade ou homogeneidade.

A análise de *cluster* designada por *fuzzy* (difusa) é uma generalização da ideia de uma partição. A ideia de partição enuncia que um elemento pertence a um e um só *cluster* sendo

efectuada esta divisão sem dúvidas. Ora nem sempre isto se verifica pois ocorrem situações em que a decisão de colocar um objecto num grupo *A* ou *B* não é assim tão linear e é rodeada de ambiguidades.

Os métodos *fuzzy* associam a cada objecto um vector cujas componentes representam o grau de ligação do objecto a cada um dos grupos *fuzzy*. Em consequência, cada grupo fica identificado por um vector de coeficientes que representam o grau de pertença de cada um dos objectos a esse mesmo grupo. Este método tem a vantagem de fornecer mais informação sobre a estrutura de dados do que os métodos hierárquicos ou de partição. Por outro lado o método exige algoritmos mais complicados, tornando-se dispendioso em termos de tempo de cálculo e havendo muitos objectos o *output* que produz é geralmente volumoso, revelando-se por vezes de difícil interpretação.

Os **métodos de sobreposição**, ao contrário de produzir *clusters* disjuntos, ou seja, que não se sobrepõem ou intersectam, consideram que em certas situações é mais significativo permitir que um objecto pertença a mais que um grupo em simultâneo. Por exemplo, em termos de interpretação de uma palavra, esta poderá ter significados diferentes.

É de salientar que este método não poderá ser interpretado da mesma forma que o método *fuzzy*, pois afirmar que existem dúvidas sobre o *cluster* a que pertence o objecto é diferente de afirmar que o objecto pertence a mais do que um *cluster* em simultâneo.

Em suma, a análise de *cluster* tem sido descrita como uma ferramenta de "descoberta" porque possui o potencial de revelar relações previamente não detectadas. Algumas aplicações da análise de *cluster* consistiam em determinar relacionamentos taxionómicos entre espécies, perfis psiquiátricos, imagens e propriedades químicas, entre muitas outras.

6. Modelos de Mistura

Os modelos de mistura finita são usados há mais de 100 anos, no entanto, o verdadeiro impulsionador da sua utilização foi o desenvolvimento a nível informático nos últimos anos. Consideremos que os dados foram gerados através de um processo hierárquico, ou seja, cada observação foi gerada numa primeira instância a partir de um espaço discreto e finito de *clusters*, e condicional à observação do *cluster* a observação específica é gerada. Pelo facto desta abordagem ser baseada num modelo probabilístico e paramétrico as vantagens

principais da sua utilização com fins de *clustering* são: permitir modelar dados de diferentes naturezas e avaliar hipóteses sobre o modelo.

Seja $x = (x_1, \dots, x_n)$ uma amostra J – dimensional de tamanho n . Seja $\{(Z_1, X_1), \dots, (Z_n, X_n)\} = \{(Z_i, X_i)\}_{i=1}^n$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas, assumindo valores em $\nabla \times N^J$, com $\nabla = \{1, 2, \dots, S\}$ e $N \subseteq IR$. A variável discreta não observada $Z_i \in \nabla$ indica a componente que gerou a observação i , e $z = (z_1, \dots, z_n)$ representa os valores amostrais. O problema de inferência consiste em estimar os parâmetros do modelo, φ , quando apenas a sequência x é observada, uma vez que a sequência $z = (z_1, \dots, z_n)$ é latente.

Em consequência, qualquer método de estimação terá de se basear na distribuição marginal de x :

$$\begin{aligned} f(x_i | \varphi) &= \sum_{s=1}^S p(Z_i = s, X_i = x_i | \varphi) = \\ &= \sum_{s=1}^S p(Z_i = s | \varphi) p(X_i = x_i | \varphi) = \\ &= \sum_{s=1}^S w_s f_s(x_i | \theta_s) \end{aligned} \quad (7)$$

que define um modelo de mistura finita com S pontos de suporte ou componentes. As proporções da mistura, $w_s = p(Z_i = s | \varphi)$, correspondem à probabilidade *a priori* do indivíduo i pertencer à componente s . Esta distribuição da mistura, $\{w_s\}_{s=1}^S$, uma vez que é uma função de probabilidade, satisfaz as condições:

$$\begin{aligned} w_s &> 0 \\ \sum_{s=1}^S w_s &= 1 \end{aligned} \quad (8)$$

Em cada componente, a observação x_i é caracterizada pela densidade $f_s(x_i | \theta_s) = p(X_i = x_i | Z_i = s; \varphi)$. A função $f_s(x_i | \theta_s)$ implica que todos os indivíduos pertencentes a uma das componentes têm a mesma distribuição de probabilidade, apenas os parâmetros θ_s variam entre as várias componentes. Os parâmetros do modelo de mistura finita a serem estimados são $\varphi = (w_1, \dots, w_{S-1}, \theta_1, \dots, \theta_S)$.

O objectivo nestes problemas é, em simultâneo, determinar o número de componentes s e o valor de $f_s(x_i | \theta_s)$ para cada objecto e classificador. Sabendo $f_s(x_i | \theta_s)$ pode-se adoptar a regra do grupo modal que atribui cada objecto a um dos grupos definidos.

Capítulo II – Aspectos Genéricos das Árvores de Classificação

As Árvores de Classificação/Decisão são uma técnica utilizada na construção de modelos de análise de dados e na sua classificação. Neste capítulo veremos o que são, qual o seu interesse e como construí-las.

1. Árvores de Decisão / Classificação: Fundamentos

Uma das principais características de uma Árvore de Decisão é o seu tipo de representação: uma estrutura hierárquica que traduz uma árvore invertida que se desenvolve da raiz para as folhas. A representação hierárquica traduz uma progressão da análise de dados no sentido de desempenhar uma tarefa de previsão/classificação. Em cada nível da árvore tomam-se decisões acerca da estrutura do nível seguinte até atingir os nós terminais (nós folha).

O princípio subjacente à utilização deste tipo de modelos é o princípio de dividir-para-conquistar. Deste modo, em cada nível de uma árvore, um problema mais complexo de previsão/classificação (em que há maior heterogeneidade de valores da variável alvo) é decomposto em subproblemas mais simples. Isto traduz-se na geração de nós descendentes, nos quais, a heterogeneidade da variável a prever (e explicar) é mais atenuada, podendo as previsões serem efectuadas com menos riscos, para cada um desses nós. Trata-se de uma pesquisa que se desenvolve do geral para o particular, no sentido em que cada novo nível de nós descendentes se limita (particulariza) o valor de mais um atributo explicativo.

Assim sendo, podemos definir Árvore de Decisão/Classificação como uma estrutura de dados recursivamente definida com nós folha, que indicam uma classe, ou nós de decisão que contém um teste sobre o valor de um atributo. Para cada um dos possíveis valores do atributo, tem-se um ramo para uma outra árvore de decisão (sub-árvore). Cada sub-árvore contém a mesma estrutura de uma árvore. Árvores de Decisão dividem o espaço de descrição do problema em regiões disjuntas, isto é, um exemplo é classificado por apenas um único ramo da árvore. É um método de classificação supervisionado, onde uma variável dependente é explicada à custa de n variáveis independentes medidas em qualquer escala.

As Árvores de Decisão/Classificação podem ser usadas com objectivos diferentes, de acordo com o problema que se pretende resolver. Podemos ter por objectivo classificar os dados referentes a uma população da forma mais eficiente possível ou descobrir qual é a estrutura de um determinado tipo de problema, compreender quais as variáveis que afectam a sua resolução e construir um modelo que o solucione. Com uma Árvore de Decisão/Classificação é possível escolher as variáveis explicativas que realmente nos interessam para descrever a situação, deixando de lado as menos relevantes.

Podem identificar-se algumas vantagens na utilização de Árvores de Decisão/Classificação, das quais se destacam as seguintes:

7. Ausência dos pressupostos típicos de modelos paramétricos de verificação difícil, especialmente se o número de variáveis explicativas é elevado;
8. Possibilidade de utilização de variáveis explicativas em qualquer número e em várias escalas de medida e disponibilização de técnicas para lidar com valores omissos, permitindo evitar extensos e demorados tratamentos prévios aos dados;
9. As variáveis podem ser utilizadas sem necessidade de transformação ou codificação como acontece com os atributos nominais em modelos de regressão ou discriminantes, onde a logaritmização por vezes é necessária para evitar problemas de heterocedasticidade;
10. Possibilidade de integração de relações complexas entre as variáveis explicativas e a dependente e não apenas relações lineares, como acontece na maioria dos procedimentos estatísticos;
11. Interpretabilidade dos resultados muito simples e clara, por simples observação da árvore.

As principais desvantagens centram-se essencialmente:

12. Na instabilidade, pois pequenas perturbações do conjunto de treino podem provocar grandes alterações no modelo aprendido;
13. Na fragmentação de conceito, ou seja, podem ocorrer replicações de sub-árvores.

Esta é uma técnica que pode ser utilizada em diversos campos. Por exemplo na saúde, combinando informação recolhida através de inquéritos referentes a dados clínicos para descobrir quais as variáveis que contribuem para a melhoria do sistema de saúde; na análise de mercados, verificando quais as variáveis associadas com o volume de vendas (preço, geografia, características dos consumidores, *etc.*), entre outros campos da ciência.

Podem ainda ser usadas para identificar as pessoas que pertencem a um determinado grupo alvo, ou ainda para detectar relações entre variáveis que surjam apenas num sub-grupo, especificando-as num modelo formal.

2. Crescer Árvores de Decisão / Classificação

Uma das formas de obtermos uma Árvore de Classificação/Decisão é considerarmos um conjunto de dados D , em que, $D = \{x_1, x_2, \dots, x_n\}$ e considerarmos igualmente as classes definidas por C , em que, C está definida da seguinte forma, $C = \{c_1, c_2, \dots, c_m\}$ com $m < n$. O objectivo será estabelecer uma relação segundo uma função f definida como $f : D \rightarrow C$, em que, a cada vector x de D (o espaço de medida), cujas coordenadas são os valores assumidos pelas variáveis explicativas que descrevem a amostra para o caso x , a uma classe de C . Podemos definir então a regra de classificação ou classificador como *"uma função $f(x)$ definida em D , que para todo o $x \in D$, $f(x)$ pertence a uma das classes c_1, c_2, \dots, c_m , ou equivalentemente, "uma partição de D em m subconjuntos disjuntos A_1, \dots, A_K , com:*

$$D = \bigcup_{j \in C} A_j \tag{9}$$

tal que, para todo o $x \in A_j$ a classe prevista é a j ".

Se a amostra de treino for perfeitamente representativa da população (caso raro) e a árvore tiver a dimensão correcta, os erros cometidos sobre a amostra (proporção de casos incorrectamente classificados) são boas estimativas dos erros de previsão da população em geral.

No entanto, na prática, raramente temos disponíveis os dados referentes à população, ou outras amostras de tamanho suficientemente grande, os dados de D têm de ser utilizados para construir $f(x)$ e para estimar o valor do risco de classificação incorrecta, $R^*(f)$, associado a f . A este tipo de validação chamamos **validação interna**. Consideremos o método de

validação *Jackknife*, embora existam diversas formas de a realizar. O *Jackknife* é um método não paramétrico destinado a estimar o enviesamento (e, portanto, reduzi-lo) e a variância de estimadores em condições teoricamente complexas ou em que não se tem confiança no modelo especificado. É um método de reamostragem pois baseia-se na construção de sub-amostras da amostra inicial. Desta forma, a amostra original é particionada em v sub-amostras, de dimensões semelhantes, com uma distribuição associada semelhante à da variável dependente. Numa primeira etapa, uma das amostras é reservada para o cálculo do erro de classificação (amostra *hold-out*), enquanto as restantes $v-1$ amostras são utilizadas como base para a construção das regras de classificação. Este processo repete-se até que cada uma das v amostras tenha sido utilizada como amostra *hold-out*. O valor final do erro de classificação é obtido através da média associada às v medidas de precisão disponíveis.

3. Um exemplo de uma Árvore de Classificação

A Figura 3 apresenta um exemplo de Árvore de Classificação. Este exemplo considera objectos que relatam as condições propícias de uma pessoa ter a possibilidade ou não de contrair um empréstimo. É considerada a probabilidade do montante do empréstimo ser médio, baixo ou alto.

Alguns objectos são exemplos positivos de uma classe “sim”, ou seja, os requisitos exigidos a uma pessoa, por um banco, são “satisfatórios à concessão de um empréstimo”, e outros são negativos, onde os requisitos exigidos “não são satisfatórios à concessão de um empréstimo” – classe “não”. Os dados para fazer crescer a árvore são dados que incluem o montante do empréstimo; salário e conta, para cada indivíduo, sem esquecer uma variável indicando se o empréstimo foi concedido ou não, na amostra de treino.

Classificação, neste caso, é a construção de uma estrutura de árvore, que pode ser usada para classificar os novos objectos do conjunto com um mínimo de erro.

A partir de uma Árvore de Decisão é possível derivar regras. As regras são escritas considerando o caminho do nó raiz até uma folha da árvore. A derivação de regras e a construção da árvore de decisão são geralmente utilizados em conjunto. Uma das principais razões para esta utilização é o facto das árvores de decisão tenderem a crescer muito, daí muitas vezes, serem substituídas por regras. Isto acontece em virtude das regras poderem ser

facilmente modularizadas. Em contrapartida, a geração de um conjunto de regras é, geralmente, mais morosa que a construção de uma árvore de decisão. Uma regra pode ser compreendida sem que haja a necessidade de referenciar-se outras regras.

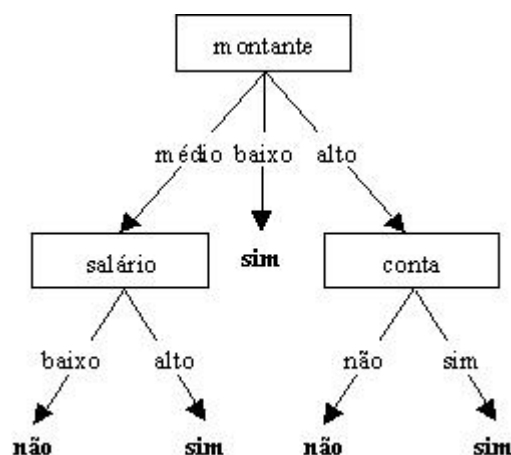


Figura 4 – Exemplo de uma árvore de classificação.

É de salientar que embora uma árvore de decisão seja perfeitamente traduzível num conjunto de regras proposicionais, o oposto nem sempre se verifica, pois um conjunto de regras não resulta, necessariamente, de um processo hierárquico como o associado à construção de Árvores de Decisão.

Com base na Árvore de Decisão/Classificação apresentada na figura 3, pode-se exemplificar a derivação de regras. Dois exemplos de regras obtidas a partir desta árvore são os seguintes:

- Se montante = médio e salário = baixo
então classe = não
- Se montante = médio e salário = alto
então classe = sim

Após a construção de uma Árvore de Decisão é importante avaliá-la. Esta avaliação é realizada através da utilização de dados que não tenham sido usados no treino. Esta estratégia permite estimar como a árvore generaliza os dados e se adapta a novas situações, podendo também, estimar a proporção de erro e acertos ocorridos na construção da árvore.

Vários são os problemas a serem superados em qualquer algoritmo de construção de Árvores de Decisão/Classificação para que esta seja óptima em aspectos como dimensão, erro

resultante da classificação, tempo de construção, entre outros. Alguns destes, que ainda hoje são tema de pesquisa, estão abaixo referidos:

14. Escolha da melhor partição para um nó - em geral, por escolha do atributo ou variável e pontos de corte;
15. Estratégias para limitação no crescimento da árvore;
16. Tratamento de valores omissos no conjunto de objectos para treino e para teste;
17. Custo de classificações erradas;
18. Integração de conhecimento prévio.

Capítulo III – Principais Algoritmos de Árvores de Classificação e Decisão

Constituindo-se como os pioneiros em técnicas não paramétricas de aprendizagem supervisionada, Morgan e Sonquist (1963) apresentam aplicações especialmente adequadas para grandes volumes de observações e variáveis explicativas em várias escalas de medida. O método utilizado por estes autores é conhecido por *AID – Automatic Iteration Detector* e baseia-se na análise de variância para segmentar as observações em grupos distintos para os quais podem ser desenvolvidos modelos de previsão causais.

Os principais algoritmos de árvores de decisão que iremos abordar são, evolução deste primeiro algoritmo: o *ID3 – Iterative Dichotomizer 3* (Quinlan, 1986), o *C4.5* – (Quinlan, 1993), o *CHAID – Chi-square Automatic Interaction Detection* (Kass, 1980) o *CART – Classification and Regression Trees* (Breiman et al., 1984), e o *QUEST – Quick, Unbiased, Efficient Statistical Tree* (Loh e Shih, 1997).

O método adoptado por estes algoritmos consiste na divisão recursiva do conjunto de observações em subgrupos filhos construindo uma árvore da raiz para as folhas. Em cada passo, o algoritmo determina uma regra de classificação, seleccionando uma variável e um ponto de corte nos valores dessa variável, que maximize uma medida de entropia dos nós “filhos” relativamente ao nó “pai” (*C4.5* e *ID3*), minimize uma medida de impureza (*CART*), ou que maximize a distinção estatística dos “filhos” relativamente à variável dependente (*CHAID* e *QUEST*).

O principal objectivo é obter divisões dos dados que permitam definir grupos homogéneos, relativamente à variável dependente. Este processo é caracterizado pela sua repetição até que uma regra de paragem seja atingida, a qual pode ser a incapacidade de encontrar novas variáveis que permitam divisões dos dados estatisticamente significativas ou um nível máximo de dimensão da árvore.

Alguns algoritmos, como o *CART* ou *C4.5*, permitem ainda um método designado por **poda da árvore**. Muitas vezes obtemos uma árvore que modela os dados de uma forma excessivamente precisa, ou seja, é demasiado grande ou subdivide excessivamente o conjunto de dados, nesse caso, dizemos que o modelo está sobreajustado. Neste caso o modelo obtém um desempenho quase perfeito nos dados usados para estimação do modelo, mas um desempenho pobre em novos dados.

Torna-se assim necessário evitar esta característica indesejável. Uma forma de o fazermos é através da poda da árvore de decisão/classificação, isto é, remoção ou corte de ramos e sub-árvores que não nos interessam para a resolução do problema, ou que não têm interesse prático, obtendo uma árvore podada.

Existem duas formas de proceder à poda de uma árvore: não a deixando crescer até ao fim (pré-poda) ou então, cortando ramos depois de estar completa (pós-poda).

Para seleccionarmos quais os ramos a podar, calculamos o custo de classificação incorrecta, $R^*(d)$, da árvore completa e depois de o ramo ser retirado. Se este aumentar, não podamos esse ramo; se permanecer inalterado ou tiver uma variação considerada muito reduzida, o ramo em questão é dispensável. Com a poda obtemos várias árvores alternativas para modelar o problema. De forma simples podemos dizer que a selecção da melhor será realizada através da remoção dos ramos que menos contribuem para a diminuição do erro de previsão.

Estas técnicas de análise de dados são direccionadas para volumes de dados consideráveis. Assim, a qualidade dos resultados está associada a factores como o número de observações; número de variáveis disponíveis; graus de liberdade e às técnicas de amostragem utilizadas. Desta forma, o número elevado de observações e variáveis explicativas necessário pode constituir uma das principais desvantagens da utilização destas técnicas.

1. Algoritmos ID3 e C4.5

A sigla ID3 *significa Iterative Dichotomizer 3* e foi um método desenvolvido por Quinlan (1986). O algoritmo ID3 consiste num processo de indução de árvores de decisão. A construção da árvore é realizada de cima para baixo (*top-down*), com o objectivo de escolher sempre o melhor atributo para cada nó de decisão da árvore. É um processo recursivo que após ter escolhido um atributo para um nó, começando pela raiz, aplica o mesmo algoritmo aos descendentes desse nó, até que certos critérios de paragem sejam verificados.

A escolha do atributo de partição é concretizada tendo em conta o ganho de informação. O **Ganho de Informação** é uma medida estatística que está por base na construção de árvores de decisão/classificação neste algoritmo. Esta medida estatística consiste no seguinte:

Se tivermos um conjunto de vários exemplos S , e um conjunto de n classes $C = \{C_1, C_2, \dots, C_n\}$, sendo p_i a probabilidade da classe C_i em S , então a entropia do conjunto S , é a homogeneidade deste, traduzida na seguinte igualdade:

$$Entropia = - \sum_{i=1}^c p_i \log_2 p_i \quad (10)$$

A entropia é uma medida aplicável à partição de um espaço de probabilidade, medindo quanto esse espaço é homogêneo, ou por outro lado, quanto maior a entropia maior a desordem. A entropia atinge o seu valor máximo, igual a $\log_2 n$, quando $p_1 = p_2 = \dots = p_n = 1/n$, expressando precisamente a existência de uma máxima de heterogeneidade. Pelo contrário a homogeneidade máxima corresponderia a $p_1 = p_2 = \dots = p_n = 0$ e $p_i = 1$.

De outro modo, pretende-se saber qual o ganho de informação do atributo A , que é dado pela seguinte fórmula:

$$Ganho(S, A) = Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v) \quad (11)$$

em que, $valores(A)$ é o conjunto de todos os valores possíveis para o atributo A , e $|S_v|$ é o subconjunto de S para o qual o atributo A tem valor v :

$$S_v = \{s \in S \mid A = v\} \quad (12)$$

Desta forma, o Ganho de Informação, mede a eficácia de um atributo em classificar os dados de treino, a escolha do atributo mais eficaz – que mais reduz a entropia – faz com que a tendência seja a de gerar árvores, que são, em geral, menos profundas com menos nós e ramificações.

Em suma, o algoritmo *ID3* realiza uma procura ávida (*greedy*) no espaço das árvores de decisão, consistentes com os dados, guiada pelo ganho de informação e feita segundo a estratégia do “subir a colina” (*hill-climbing*). No entanto, no uso desta estratégia corre-se o risco da solução convergir para um ótimo local.

Para os atributos cujos domínios sejam valores quantitativos, reordenam-se as instâncias, de acordo com esse atributo e procuram-se pontos extremos nos quais existe uma mudança de valor da classe. Um ponto de mudança de classe marca uma partição binária do conjunto das instâncias, mediante uma condição lógica do tipo $A > x$, sendo A o atributo numérico em causa e x um valor calculado a partir dos dois valores consecutivos de A nesses

pontos. Normalmente, toma-se x igual à média dos valores de A , nos pontos consecutivos. Foi mostrado que, neste tipo de atributos, de todos os possíveis pontos de partição, aqueles que maximizam o ganho de informação correspondem exactamente à separação dos dois exemplos pertencentes a classes diferentes.

Uma das grandes vantagens do *ID3* é a sua simplicidade, o seu processo de construção torna relativamente simples a compreensão do seu funcionamento.

A maior desvantagem do *ID3* é que a árvore de decisão produzida é essencialmente imutável – não se pode eficientemente reutilizar a árvore sem a reconstruir. Usando este algoritmo para actualização, o método tende a produzir uma árvore de decisão que está longe da árvore de decisão óptima, impedindo assim a ideia original de reformular a árvore de decisão a partir da original.

O **algoritmo C4.5** (Quinlan, 1993) é um método melhorado relativamente ao *ID3* que, entre outros melhoramentos, combate o problema de *overfitting*, utilizando uma estratégia de poda de árvore. Existem duas estratégias de combate ao problema do sobreajustamento, que pressupõe que a árvore tem uma complexidade inadequada.

O princípio orientador deste algoritmo é o denominado princípio de *Occam* ou *Occam's razor*, criado por William Occam, que dá primazia à escolha de hipóteses menos complexas, compatíveis com a realidade observada, semelhante ao conceito de parcimónia da estatística.

O algoritmo *C4.5* adopta a estratégia (pós-poda). Podar uma árvore, neste contexto, significa reduzir algumas sub-árvores a folhas, ou de outra forma, um ramo da árvore, a partir de determinado nó é cortado (transformado em folha). O corte dum ramo da árvore é guiado por um teste estatístico que tem em conta os erros num nó e a soma dos erros nos nós que descendem desse nó. Assim, para cada nó, a poda só se concretiza se o desempenho da árvore não diminuir significativamente. Para além do problema do *overfitting*, o *C4.5* inclui soluções para problemas concretos e comuns do mundo real como: atributos com valores quantitativos; valores omissos e dados contendo ruído.

Uma outra possibilidade disponibilizada por este sistema é a capacidade de realizar validação cruzada com dois ou mais grupos (*v-fold* ou validação *Jackknife*), melhorando assim a estimativa do erro cometido pelo classificador.

Uma última característica que merece ser destacada é a possibilidade deste sistema em gerar regras de decisão a partir de árvores, e de as comparar entre si independentemente das árvores construídas.

2. Algoritmo CART

CART significa *Classification and Regression Trees* (Breiman *et al.*, 1984). Uma árvore de classificação utilizando a metodologia *CART* traduz o resultado de uma partição binária recursiva dos dados base da modelação.

Este algoritmo é um modelo de regressão não-paramétrico que estabelece uma relação entre as variáveis independentes (x), com uma única variável dependente, ou resposta, (*target*) ou alvo. O modelo é ajustado mediante sucessivas divisões binárias no conjunto de dados, para tornar os sub-conjuntos de dados da variável resposta cada vez mais homogêneos.

Para obter o número de divisões possíveis, este algoritmo considera, quando a variável explicativa tem k valores ordenados, $k-1$ divisões possíveis, quando a variável explicativa é nominal, com k categorias, consideram-se $2^{k-1}-1$ divisões possíveis.

Para seleccionar a melhor partição dos dados, procura-se minimizar a impureza dos nós folha resultantes, para isso são utilizadas medidas de impureza. O método *CART* considera três critérios possíveis para seleccionar a melhor partição de dados: Entropia (mencionada no algoritmo *ID3*), o critério de Gini e o critério de Twoing.

Cada uma destas opções pode ser adoptada em conjunto com a estrutura de custo de classificações incorrectas. Segundo o critério de Gini o grau de impureza num dado nó é dada por:

$$G(N) = 1 - \sum_{I=1}^L p^2(I | N) \quad (13)$$

Este critério é definido para uma variável nominal com L categorias, onde $p(I | N)$ é a probabilidade *a priori* da classe I se formar no nó N . Cada variável pode ser usada diversas vezes ao longo do processo de crescimento da árvore. Deste modo, este índice contabiliza a proporção de observações em cada classe da variável dependente num nó relativamente ao total, isto é, ao nó raiz. O critério de Gini assume o seu valor mínimo quando num nó correspondente a uma partição da variável dependente, apenas existem observações

pertencentes a uma classe. A diferença entre o critério de Gini para o nó pai e a soma dos valores para os nós filhos (ponderada pela proporção de casos em cada filho) é apresentada na árvore como *improvement*. A variável explicativa escolhida é aquela que garante uma partição correspondente ao maior valor de *improvement*.

O critério de Twoing consiste em separar a variável alvo em duas “grandes classes”, e deste modo encontrar a melhor partição na variável explicativa com base na divisão efectuada em duas classes. A função do critério de Twoing para separar s no nó t é definida como:

$$\phi(s, t) = \frac{p_L p_R}{4} \left[\sum_j |p(j | t_L) - p(j | t_R)| \right]^2 \quad (14)$$

onde t_L e t_R são os nós criados pela divisão s . A divisão s é escolhida como a divisão que maximiza este critério. Este valor, pesado pelo número de todos os casos no nó t , é o valor designado como *improvement* na árvore. Estas “grandes classes” C_1 e C_2 são definidas como:

$$\begin{aligned} C_1 &= \{j : p(j | t_L) \geq p(j | t_R)\} \\ C_2 &= C - C_1 \end{aligned} \quad (15)$$

onde C é o número de categorias da variável alvo.

As árvores obtidas através do algoritmo *CART* têm, normalmente, muitos níveis, o que pode tornar pouco eficiente a apresentação dos resultados e tornar as conclusões obtidas a partir da sua estrutura pouco fiáveis. Este algoritmo, apesar de flexível, é complexo tornando o cálculo dos resultados muito demorado para grandes conjuntos de dados.

As principais vantagens do Algoritmo de *CART* são:

- Poder utilizar variáveis independentes de diferentes tipos, desde contínuas, ordinais e nominais.
- Não obrigar à realização de transformações das variáveis iniciais independentes (como a logaritmização ou normalização) pois o método tem bom comportamento para qualquer tipo de dados.
- Poder usar a mesma variável em diferentes estágios do modelo, permitindo reconhecer efeitos que certas variáveis produzem sobre outras.

- Não necessitar de satisfazer qualquer condição de aplicabilidade do modelo, o que não acontece nos modelos paramétricos.

4. Algoritmo *CHAID*

É um dos métodos mais antigos de Árvore de Classificação proposto originalmente em 1978, por Kass (1980). O algoritmo *CHAID* - *Chi-squared Automatic Interaction Detector* tem por base os testes de Qui-Quadrado de Pearson numa tabela de contingência entre as categorias da variável dependente e as categorias das variáveis independentes (as variáveis contínuas são previamente discretizadas em classes). Constitui um método estatístico extremamente eficiente para a segmentação, ou crescimento de uma árvore.

De facto, realiza-se um conjunto de testes agregando as classes da variável explicativa até restarem apenas duas, de modo a descobrir o melhor número de classes. Este processo repete-se para a totalidade das variáveis explicativas e a melhor variável explicativa com o melhor conjunto, isto é, a menor probabilidade de significância (*p value*) ajustada pelo método Bonferroni, é escolhido.

O teste estatístico usado neste algoritmo depende da característica da variável alvo. Se a variável alvo é contínua, o teste Fisher é implementado, se a variável alvo é nominal, então o teste do Qui-Quadrado de Pearson é implementado, se a variável alvo é ordinal o teste de rácio de verosimilhança (*likelihood-ratio*) é implementado.

Uma das vantagens deste algoritmo, é o facto de parar o crescimento da árvore antes do problema de *overfitting* ocorrer, isto é, não tem tratamentos como o da poda. Uma das desvantagens do algoritmo *CHAID* é o facto de requerer grandes quantidades de dados para ser possível assegurar que a quantidade de observações dos nós folha é significativa.

5. Algoritmo QUEST

O algoritmo *QUEST* foi criado em 1997 e significa, *Quick, Unbiased, Efficient Statistical Tree*, e como o nome indica coloca claramente a tónica na eficiência. As árvores obtidas estão sujeitas a um menor enviesamento e o tempo de cálculo é mais reduzido que nos outros dois algoritmos.

Da mesma forma que o algoritmo *CART*, também é binário. No entanto, separa o processo de selecção das variáveis de partição do processo de busca da melhor segmentação dos dados em classes. Pode ser aplicado a qualquer tipo de variáveis preditivas ou explicativas, mas a variável dependente tem de ser nominal. Se várias das variáveis preditivas ou explicativas possuem o mesmo valor informativo, então todas têm a mesma probabilidade de ser escolhidas.

Este método utiliza, tal como o algoritmo *CHAID*, testes de Qui-Quadrado de Pearson para tabelas de contingência. No entanto, utiliza um maior conjunto de testes estatísticos para garantir a independência entre o processo de selecção da variável explicativa e o ponto de divisão das classes da mesma variável. Usa, por exemplo, a estatística *F* de *ANOVA* e a estatística *F* de Levene para variâncias não iguais na selecção de variáveis métricas. Utiliza-se o método das 2-médias de análise de *clusters* para agrupar as classes da variável dependente e análise discriminante quadrática na escolha do ponto de divisão da variável explicativa. Assim, nas árvores sempre binárias, apresentam-se valores para o teste de Qui-Quadrado ou estatística *F* de *ANOVA* ou Levene (Loh e Shih, 1997).

As vantagens deste algoritmo são determinadas pelas suas principais características: rapidez e eficiência. Em contrapartida, este algoritmo apresenta também desvantagens, sendo a mais relevante o facto da variável dependente só poder ser nominal, pelo que não é possível construir árvores com variáveis dependentes contínuas.

Capítulo IV – Caso de Estudo

Este capítulo apresenta a aplicação dos modelos apresentados no capítulo III em relação a uma amostra de dados em concreto. Os dados foram recolhidos através de vários dias de observação a uma área extremamente problemática em termos operacionais, em qualquer aeroporto do país. A zona do *check-in* por vários factores constitui o primeiro grande filtro no processo de embarque de um passageiro. Desta forma, o estudo foi elaborado com uma amostra de 10 voos “tipo” do aeroporto João Paulo II em Ponta Delgada. Considera-se o voo que se realiza diariamente para Lisboa às 21:25, e o dia de maior movimento quer de passageiros quer de aviões, que neste caso é a Segunda-Feira.

Assim sendo foram considerados intervalos de 15 em 15 minutos, em que foi contabilizado o número de passageiros por fila de *check-in*. Foi tido em conta a hora real de chegada do avião, por ter influência na chegada dos passageiros à aerogare, a hora real de partida pois em todos os voos acontecem ligeiras, e por vezes, grandes oscilações à hora definida. Outro aspecto importante a ter em conta, foi o equipamento utilizado para a realização dos voos, por exemplo, o Airbus 310 da SATA Internacional tem capacidade máxima de 220 passageiros, enquanto o Airbus 320 tem capacidade máxima de 161 passageiros, o qual pode ser um factor importante na fila a criar.

Igualmente, foi tido em conta o número de passageiros embarcados por voo. Um factor de influência no número de passageiros por fila, são também o número de balcões de *check-in* atribuídos aos voos. A forma como estes são atribuídos pode ser problemático conduzindo à aglomeração de passageiros em determinados balcões pois poderão não dar a passagem a outros passageiros para realizar o seu *check-in*. Um aspecto de relevo também foi considerar uma variável que define critérios de qualidade, em que, é considerável aceitável uma fila máxima de 9 passageiros por *check-in* de 15 em 15 minutos, caso contrário a fila é considerada não aceitável, em termos de tempo de espera do passageiro.

Todas estas variáveis foram utilizadas de forma a aplicar os algoritmos apresentados no capítulo III, usando o programa Answer Tree 2.0.

O Algoritmo utilizado no primeiro exemplo foi o Algoritmo CART.

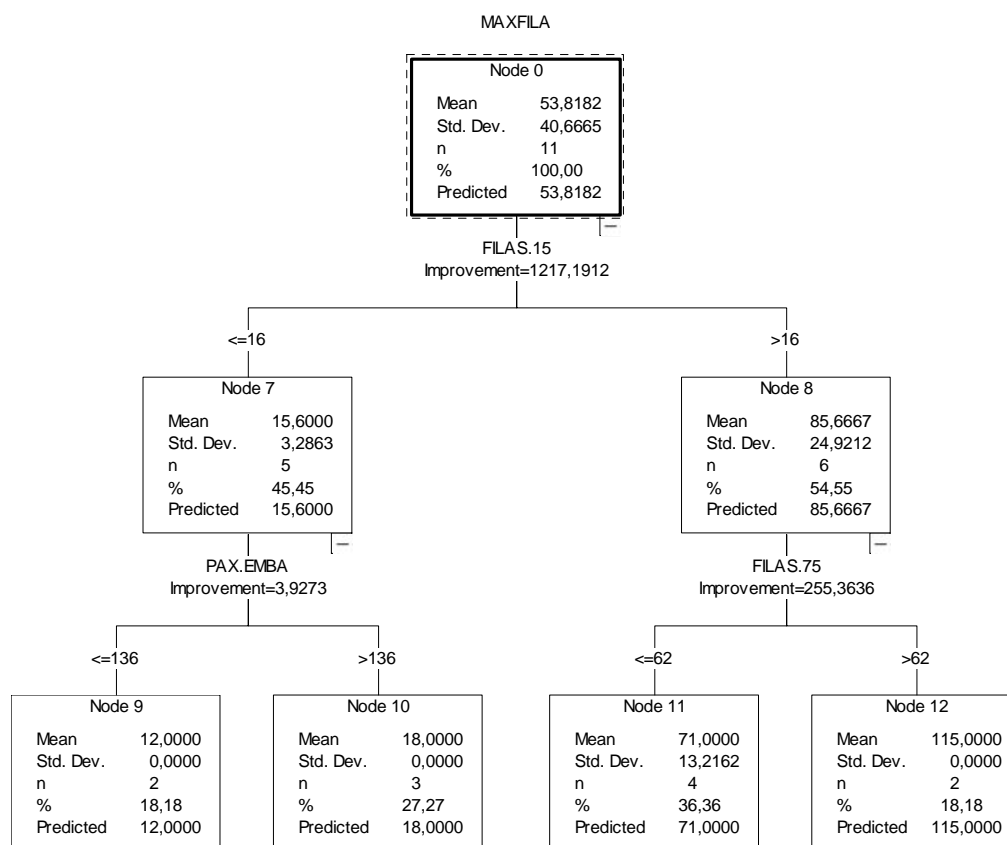


Figura 5 – Árvore de Regressão pelo Método CART.

A variável dependente escolhida foi o “número máximo de passageiros por fila”, e entre as variáveis referidas acima, o algoritmo seleccionou como melhores variáveis preditivas as “filas com os primeiros 15 minutos de *check-in* aberto”, “os passageiros embarcados” e “as filas no *check-in* com 75 minutos de funcionamento”.

No primeiro nó, obtemos a descrição do número máximo de passageiros da amostra relativamente ao *check-in* utilizado para cada voo.

Em seguida, a amostra foi classificada de acordo com o número de passageiros nos primeiros 15 minutos de *check-in* aberto, para cada voo. Verificou-se que se as filas nos primeiros 15 minutos tem 16 ou menos passageiros então existe uma probabilidade de 45,5% de a fila vir a exceder os 15 passageiros, caso contrário, existe uma probabilidade de 54,5% de virmos a ter valores na ordem dos 85 passageiros como máximo nas filas.

As filas correspondentes a 16 ou menos passageiros foram segmentados no número de passageiros embarcados, se este for inferior ou igual a 136 passageiros prevê-se um número

médio de 12 passageiros como máximo nas filas, caso contrário, prevê-se um número máximo de 18 passageiros nas filas.

Por fim, as filas com mais de 16 passageiros nos primeiros 15 minutos, foram divididas separando o nó de passageiros ao fim de 75 minutos, se nesta altura as filas de passageiros forem inferiores a 62 passageiros então a previsão será de no máximo 71 passageiros, caso contrário as filas poderão chegar aos 115 passageiros.

No segundo exemplo consideremos o Algoritmo CHAID.

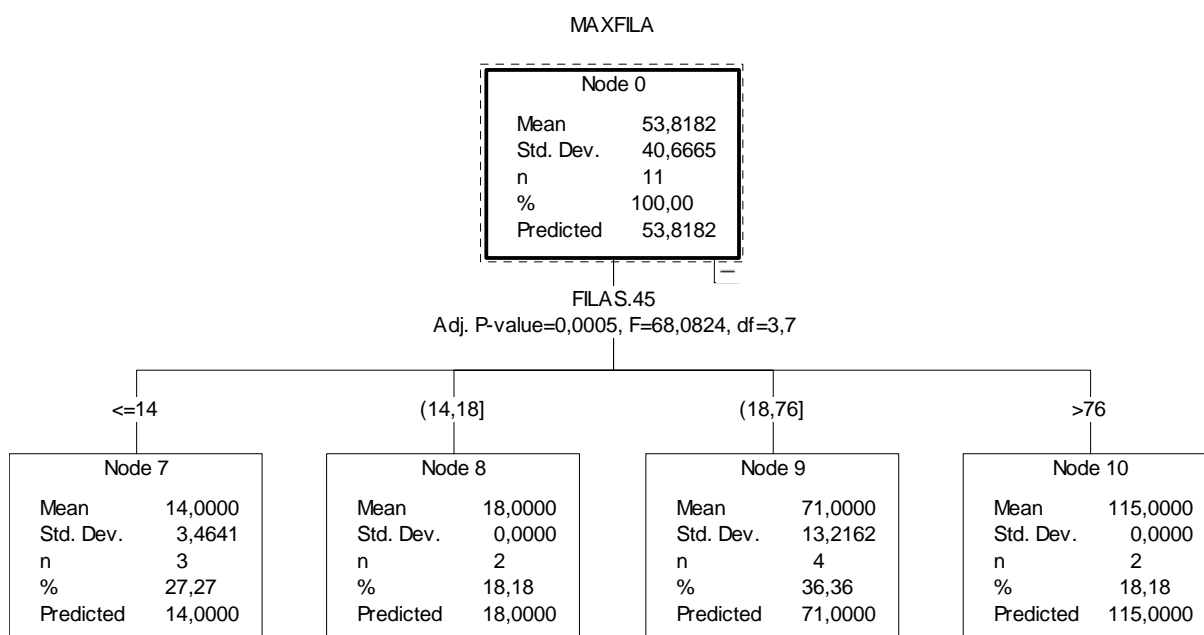


Figura 6 – Árvore de Regressão pelo Método CHAID.

A variável dependente escolhida foi, como no caso anterior, o número máximo de passageiros por fila, e das variáveis referidas acima o algoritmo seleccionou como melhor variável preditivas o comprimento das filas ao fim dos primeiros 45 minutos.

O Algoritmo faz uma previsão tendo em conta: se o comprimento das filas no *Check-In* forem inferiores ou iguais a 14 passageiros, então a previsão média do máximo de passageiros nas filas será de 14, se for entre 14 e 18 passageiros a previsão média do máximo de passageiros nas filas será de 18, se for entre 18 a 76 passageiros a previsão média do máximo de passageiros nas filas será de 71 passageiros e por último se for maior que 76 passageiros então a previsão média do máximo de passageiros será de 115.

No último exemplo consideremos o Algoritmo *QUEST*.

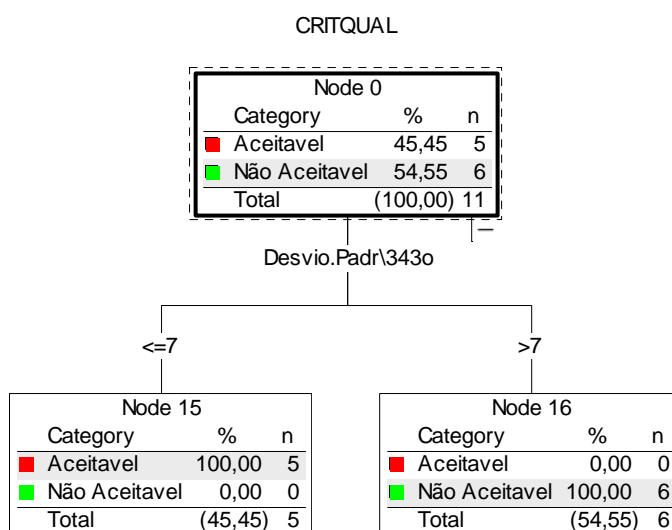


Figura 7 – Árvore de Classificação pelo Método QUEST.

A variável dependente escolhida foi o Critério de Qualidade definido como atrás, e das possíveis variáveis preditivas o algoritmo seleccionou como melhor variável do comprimento das filas o Desvio Padrão da variação.

Segundo o nosso critério de qualidade classificou-se 45,4% de voos com filas no *check-in* aceitáveis e 54,5% não aceitáveis.

O Critério de Qualidade foi segmentado pelo desvio padrão, se este verificar valores baixos em relação ao processo de chegada de passageiros ao *check-in* então existe 100% de hipóteses de satisfazer o critério de Qualidade, caso contrário, a probabilidade é de 100% para uma hipótese de não aceitável, ou seja, não satisfazer o critério de qualidade.

Conclusão

Após a realização deste trabalho: Árvores de Classificação, foi possível reconhecer o grande interesse da temática abordada para um aluno de matemática, na medida em que, para um tema à partida desconhecido, foi possível um enriquecimento dos meus conhecimentos através do estudo e pesquisa efectuados, bem como através da sua aplicação prática no mundo real.

A aplicação da técnica de Árvore de Decisão, permitiu transformar dados analisados em informações úteis e importantes para tomada de decisões no âmbito da área profissional onde estou inserido: Operações Aeroportuárias.

O trabalho prático realizado permitiu que, através de um conjunto de dados recolhidos na área do *check-in*, de alguns voos, fosse possível determinar previsões sobre o comprimento máximo das filas de passageiros no *check-in*, ou se os critérios de validade se verificam. Pese embora o facto dos resultados não poderem ser validados, dado o reduzido número de voos, objecto de recolha de dados.

Os resultados obtidos no trabalho realizado servirão como “ponto de partida” para potenciar previsões mais minuciosas sobre várias áreas operacionais relacionadas, neste caso, com a aviação. O objectivo é permitir apoiar decisões como o momento adequado para abrir novas caixas de *check-in* ou obter informação para poder realizar simulações numéricas e avaliar o desempenho de diferentes cenários de organização de todo o procedimento operacional.

Bibliografia

- BAGOZZI, Richard P. (Ed.) (1994) “Advanced Methods of Marketing Research”; Blackwell Publishers: Cambridge, USA. ISBN: 1-55786-549-3.
- BIGGS, D., B. de Ville, and E. Suen. (1991) “A method of choosing multiway partitions for classification and decision trees” *Journal of Applied Statistics*, 18: 49-62
- BRANCO, João A. (2004) “Uma Introdução à Análise de Clusters”; SPE: Évora, Portugal. ISBN: 972-98619-9-4.
- BREIMAN, Leo; FRIEDMAN, Jerome H.; OLSHEN, Richard A. e STONE, Charles J. (1984) “Classification and Regression Trees”; Wadsworth International: California, USA. ISBN: 0-534-98053-8.
- DUDA, Richard O.; HART, Peter E. e STORK, David G. (2001) “Pattern Classification”; Wiley-Interscience: New York, USA. ISBN: 0-4-710-5669-3.
- FERREIRA, Manuel Alberto M.; MENEZES, Rui e CATANAS, Fernando (2004) “Temas em Métodos Quantitativos” vol. 4; Sílabo: Lisboa. ISBN: 972-618-329-4.
- JAJUGA, Krysztof; SOKOLOWSKI, Andrzej; BOCK, Hans-Hermann (Eds.) (2002) “Classification, Clustering, and Data Analysis: Recent advances and applications”; Springer-Verlag: Berlin, Alemanha. ISBN: 3-540-43691-X.
- HAND, David J.; MANNILA, Heikki e SMYTH, Padhraic (2001) “Principles of Data Mining”; MIT Press: Cambridge, USA. ISBN: 0-262-08290-X.
- KASS, G (1980) “An exploratory technique for investigating large quantities of categorical data” *Applied Statistics*, 29:2, 119-127.
- LOH, Wei-Yin e SHIH, Yu-Shan (1997) “Split selection methods for classification trees” *Statistica Sinica*, vol. 7, pp. 815-840.
- MARQUES, Jorge Salvador (1999) “Reconhecimento de Padrões: Métodos estatísticos e neuronais”; IST Press: Lisboa, Portugal. ISBN: 972-8469-08-X.
- MONTANARI, Angela e LIZZANI, Laura (2001) “A projection pursuit approach to variable selection” *Computer Statistics and Data Analysis*, vol. 35. pp. 463-473.

- MORGAN, J.N. e SONQUIST, J.A. (1963) “Problems in the analysis of survey data and a proposal”, *Journal of the American Statistical Association*, vol. 58, pp. 58-415.
- MURTEIRA, B.J.F. (1990) “Probabilidades e Estatística”, Volume II, McGraw-Hill: Portugal. ISBN: 972-9241-17-1.
- QUINLAN, J. Ross (1993) “C4.5: Programs for machine learning”; Morgan Kaufmann Publishers: San Mateo, USA. ISBN: 1-55860-238-0.
- QUINLAN, J. Ross (1986) “Introduction of decision trees”, *Machine Learning*, vol. 1, pp. 81-106.
- REIS, Elizabeth e HILL, Manuela Magalhães (Eds.) (2003) “Temas em Métodos Quantitativos”, vol. 3; Sílabo: Lisboa. ISBN: 972-618-291-1.
- REIS, E; MELO, P; ANDRADE, R; CALAPEZ, T.; 1999. *Estatística Aplicada*. Edições Sílabo Volume 1. ISBN: 972-618-195-X.
- SPSS 1998. *Answer Tree 2.0 User's Guide*. SPSS, Inc. www.spss.com.


Sites de interesse:

<http://www.stat.wisc.edu/~loh/quest.html>

<http://www.stat.wisc.edu/~loh/>

<http://name.math.univrennes1.fr/bernard.delyon/textbook/stclatre.html#comparison>

<http://www.datawarehouses.hpg.ig.com.br/>

	<p>DEPARTAMENTO DE MATEMÁTICA Secção de Estatística e Investigação Operacional Marco António dos Santos Rodrigues ©</p> <p>Rodrigues, Marco A.S. (2006) “Árvores de Classificação” <i>Monografias da SEIO</i>. Depto. Matemática da Univ. dos Açores: Ponta Delgada, www.uac.pt/~amendes (ID 54.218)</p> <p>O trabalho apresentado é da exclusiva responsabilidade do aluno que o assina. O Departamento de Matemática e a Universidade dos Açores não se responsabilizam por eventuais erros existentes no mesmo.</p> <p>Os textos podem ser descarregados livremente, impressos e utilizados para ensino ou estudo dos temas a que se referem. No entanto, não podem ser copiados ou incluídos noutros trabalhos académicos ou de qualquer outra natureza, sem o consentimento do autor e a devida referência completa. Para autorização de cópia parcial ou integral, utilize o endereço de correio electrónico: seio@notes.uac.pt</p>
---	---