EXTENSÕES DO ALGORÍTMO DE ÁRVORES DE CLASSIFICAÇÃO PARA A ANÁLISE DE DADOS CATEGORIZADOS MULTIVARIADOS UTILIZANDO COEFICIENTES DE DISSIMILARIDADE E ENTROPIA

Cesar Augusto TACONELI[1]
Silvio Sandoval ZOCCHI[2]
Carlos Tadeu dos Santos DIAS[2]

- RESUMO: A análise estatística de grandes bancos de dados requer a utilização de metodologias flexíveis, capazes de produzir resultados esclarecedores e facilmente compreensíveis frente a dificuldades como a presença de números elevados de variáveis, diferentes graus de associações entre as mesmas e dados ausentes. A utilização de árvores de classificação e regressão proporciona a modelagem de uma variável resposta, categorizada ou numérica, com base em um conjunto de covariáveis, sem esbarrar nas dificuldades mencionadas. A extensão multivariada de técnicas de classificação e regressão por árvores tem sido estudada de maneira mais acentuada para situações em que se têm múltiplas variáveis resposta numéricas. Propõe-se neste trabalho metodologias fundamentadas em árvores de classificação multivariadas aplicadas à análise de múltiplas variáveis resposta categorizadas, com base em coeficientes de dissimilaridade e entropia. Por meio de um estudo de simulação, verificou-se que os resultados produzidos pelos métodos propostos são melhores quanto maiores as entropias e correlações das variáveis sob estudo. A análise de dados de consumo de álcool e fumo dos habitantes do município de Botucatu-SP complementa o presente estudo, indicando, dentre outras coisas, o grau de escolaridade, a ocupação profissional e a possibilidade de compartilhar problemas com amigos como fatores que influenciam o perfil de consumo de álcool e fumo dos habitantes.

- PALAVRAS CHAVE: Árvores de classificação; dissimilaridade; entropia; álcool e fumo; simulação multivariada.

1 Introdução

Levantamentos e experimentos são responsáveis, muitas vezes, pela produção de dados complexos, com grande número de variáveis e elementos, tornando necessária a aplicação de análises estatísticas sofisticadas e originando, por vezes, resultados de difícil interpretação até mesmo para profissionais da área estatística. A proposta de métodos capazes de produzir resultados de fácil compreensão em tais ocasiões torna-se, então, fundamental. Nesse contexto, técnicas de classificação e regressão por árvores (Classification And Regression Trees – CART - Breiman et al., 1984; De'Ath e Fabricius,

[1] Departamento de Estatística, Universidade Federal do Paraná – UFPR, CEP 81531-990, Curitiba, PR, Brasil. E-mail: *taconeli@ufpr.br*
[2] Departamento de Ciências Exatas, Escola Superior de Agricultura Luiz de Queiroz – ESALQ, Universidade de São Paulo USP, CEP 13418-900, Piracicaba, SP, Brasil. E-mail: *sszocchi@esalq.usp.br* / *ctsdias@esalq.usp.br*

2000) surgem como uma alternativa preditiva/exploratória de grande valia, dadas a simplicidade e a versatilidade associadas.

A construção de modelos de classificação e regressão por árvores possibilita a explicação de uma variável resposta numérica (regressão) ou categorizada (classificação) por meio de um conjunto de covariáveis e de suas eventuais interações. O método CART baseia-se na execução de partições binárias sucessivas de uma amostra, com base nos resultados amostrados das covariáveis, buscando a constituição de subamostras internamente homogêneas. A classificação dessas subamostras é realizada conforme alguma medida descritiva e a predição de novos elementos, executada por meio da estrutura de classificação constituída.

Técnicas de regressão e classificação por árvores podem ser empregadas como alternativa ou complemento a procedimentos estatísticos de regressão, agrupamentos e classificação. A versatilidade de tais técnicas é notória, comprovada por suas aplicações com finalidades similares à regressão linear múltipla, regressão logística, análise de sobrevivência, análise discriminante, correlação canônica e análise de agrupamentos, dentre outros métodos estatísticos. Além disso, o CART destaca-se por sua flexibilidade, sem quaisquer restrições quanto à natureza e à distribuição das variáveis, e por sua simplicidade, tanto em relação à construção do modelo quanto à interpretação dos resultados.

A extensão do CART à análise de dados multivariados vem sendo estudada e difundida com intensidade nos últimos anos. O mérito da modelagem conjunta de múltiplas variáveis resposta consiste na obtenção de um único modelo para a análise de múltiplas variáveis, permitindo também incorporar as possíveis correlações. A construção de árvores de classificação e regressão multivariadas requer, no entanto, critérios adequados quanto à segmentação das amostras e à avaliação da qualidade do modelo. Extensões multivariadas do CART permitem analisar dados longitudinais (Segal, 1992), respostas binárias múltiplas (Zhang, 1998) e respostas quantitativas múltiplas (De'Ath, 2002; Miller e Franklin, 2002; Larsen e Speckman, 2004). Um procedimento de classificação e regressão por árvores capaz de acomodar diferentes tipos de variáveis usando GEE (*generalized estimation equations*) é proposto em Lee (2005).

Propõem-se aqui extensões do algoritmo de classificação por árvores para a análise de múltiplas variáveis resposta, baseadas em coeficientes de dissimilaridade e entropia. Essas técnicas visam permitir o estudo de múltiplas variáveis categorizadas por meio da construção de um único modelo, conservando a estrutura de correlação dos dados e tornando mais compacta a análise. Os procedimentos propostos são avaliados por meio de um estudo de simulação. Um exemplo com dados de consumo alcoólico, cigarro e maconha, produzidos por questionários aplicados a uma amostra de habitantes do município de Botucatu (SP) complementa a análise.

2 Árvores de classificação univariadas

A construção de uma árvore de classificação é iniciada com a execução de partições de uma amostra e das subamostras constituídas, sempre originando duas novas subamostras. Denomina-se *nó inicial* à amostra original, *nós intermediários* às subamostras que dão origem a novas subamostras e *nós finais* às subamostras não partidas. Denota-se por t um nó qualquer. As referidas partições baseiam-se nos

resultados amostrados das covariáveis. Seja $\{ Y_j, \mathbf{X}_j \}, j = 1,2,...,n$, uma amostra de tamanho n de uma variável resposta categorizada Y e de um vetor de covariáveis $\mathbf{X} = (X_{1j}, X_{2j}, ..., X_{pj})$ de dimensão p. Considere X_l uma variável ordenável e τ um dos resultados amostrados para X_l. Nesse caso, pode-se partir a amostra em duas, alocando os elementos a nós distintos conforme resposta positiva ou negativa à questão $"X_{lj} \leq \tau ?"$. Caso a variável X_l não seja ordenável, considere A uma categoria (ou subconjunto de categorias) de X_l. Alocam-se elementos a nós distintos conforme resposta (positiva ou negativa) à questão $"X_{lj} \in A ?"$.

Seguindo os procedimentos descritos para segmentação de amostras, devem-se considerar todas as possíveis partições proporcionadas pelas *p* covariáveis sob estudo, respeitando, no entanto, restrições quanto ao número mínimo de elementos nos nós a serem partidos ou constituídos, a fim de não comprometer a acurácea do modelo. As partições candidatas devem então ser comparadas, executando-se aquela responsável pela formação de subamostras com menores taxas de heterogeneidade. Com o objetivo de quantificar a heterogeneidade das subamostras obtidas, Breiman *et al*. (1984) definem diferentes coeficientes denominados *medidas de impureza*. Para árvores de classificação propõem, por exemplo, a medida de entropia, apresentada em Zar (1999) como medida de diversidade de Shannon.

Seja Y uma variável categorizada e $\{ Y_1 , Y_2 , ... , Y_m \}$ seu conjunto de resultados possíveis. Define-se a entropia de um nó t como $\phi(t) = -\sum_{k=1}^{m} p_t(y_k) \log_2 (p_t(y_k))$, sendo $p_t(y_k)$ a proporção de elementos alocados ao nó *t* pertencentes à classe *k*. Quando $p_t(y_k) = 0$, considera-se $p_t(y_k) \log_2 (p_t(y_k)) = 0$. Assim, tem-se que $Mín(\phi(t)) = 0$, quando $p_t(y_k) = 1$ e $p_t(y_{k'}) = 0, \forall\, k' \neq k$, situação em que todos os elementos pertencem a uma mesma categoria (heterogeneidade mínima). Além disso, tem-se que $Máx(\phi(t)) = \log_2 k$, quando $p_t(y_k) = 1/k, \forall\, k \in \{1, 2, ..., m\}$, ou seja, quando os elementos se dividem com iguais freqüências entre as categorias da variável em questão (heterogeneidade máxima).

Suponha que um nó t seja dividido em dois novos nós (t_L e t_R), segundo uma partição s. Define-se a variação de heterogeneidade ocasionada por s como

$$\Delta_\varphi(s,t) = \phi(t) - \frac{n_L}{n}\phi(t_L) - \frac{n_R}{n}\phi(t_R) \qquad (1)$$

devendo ser selecionada e executada a partição s responsável por maximizar $\Delta_\phi(s,t)$. As subamostras originadas devem ser partidas de maneira semelhante à descrita para o nó inicial, com base no critério de partição estabelecido. O procedimento é repetido sucessivamente para os nós originados até a constituição de uma árvore com reduzido número de elementos em cada nó final.

Numa segunda etapa, inicia-se a busca por um modelo parcimonioso, ou seja, uma árvore que proporcione boa redução na heterogeneidade do nó inicial mediante

constituição de um número moderado de nós. Com essa finalidade, deve-se executar o processo de poda, que consiste na obtenção de uma seqüência de árvores de tamanhos decrescentes, a partir da árvore inicialmente produzida, com base em uma função do tipo custo-complexidade (Breiman et al., 1984). Sejam T_{MAX} a maior árvore, gerada pela execução de sucessivas partições binárias dos nós originados, e $\tilde{T}$ o conjunto de nós finais para uma subárvore T qualquer de T_{MAX}. Sejam, ainda, $|\tilde{T}|$ o número de nós finais de T e $\alpha \geq 0$ uma constante real denominada parâmetro de complexidade. Breiman et al. (1984) define a seguinte medida de custo-complexidade:

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \qquad (2)$$

sendo $R(T) = \sum_{t \in \tilde{T}} \phi(t)$ o custo de má-classificação da árvore T e $\phi(t)$ alguma medida de heterogeneidade calculada em um nó t, como, por exemplo, a medida de entropia. Aumentando o valor de α a partir de zero, obtém-se uma seqüência aninhada de árvores de tamanho decrescente, cada uma delas ótima para seu tamanho (Breiman *et al*., 1984).

A comparação das árvores, dentro da seqüência aninhada, pode ser realizada por meio dos custos de má-classificação das mesmas, estimados por validação cruzada ou por meio de uma amostra teste. A produção de um gráfico de $\hat{R}(T)$ versus $|\tilde{T}|$ permite avaliar a qualidade dos modelos conforme aumenta sua complexidade e a comparação dos resultados serve como subsídio para a seleção da melhor árvore. Breiman *et al*. (1984) propõem a *regra do desvio padrão* (*1-SE Rule*), que consiste na seleção da menor árvore responsável por um custo de má-classificação estimado que esteja a menos de um desvio padrão do menor custo de má-classificação avaliado dentre as árvores da seqüência aninhada. Uma vez escolhida a árvore, a caracterização dos nós finais se dá pela classe que aparece com maior freqüência dentre os elementos que constituem cada um deles.

3 Coeficientes de similaridade e dissimilaridade

Coeficientes de similaridade (S) são freqüentemente utilizados com o objetivo de quantificar a parecença de elementos com respeito a um conjunto de atributos categorizados. Associado a eles pode-se estabelecer coeficientes de dissimilaridade, visando quantificar a disparidade entre elementos em relação ao mesmo conjunto de atributos. Grande parte dos coeficientes de similaridade assume valores no intervalo $[0,1]$, sendo que valores próximos a zero indicam similaridade baixa, enquanto valores próximos a um sugerem similaridade elevada. Nesses casos, pode-se definir um coeficiente de dissimilaridade como $D = 1 - S$, assumindo resultados no mesmo intervalo, mas com interpretação inversa à de S.

Há uma grande variedade de coeficientes de dissimilaridade para a situação em que os atributos em questão são todos binários. Cox e Cox (2001) apresentam vários desses coeficientes, destacando suas características e aplicações. Poucas são, entretanto, as alternativas disponíveis para variáveis com mais de duas categorias. Os coeficientes de dissimilaridade mais comuns para dados categorizados baseiam-se na conversão das variáveis avaliadas em vetores de variáveis binárias. Esse tipo de procedimento, no

entanto, é pouco recomendável, à medida que a transformação das variáveis originais em vetores de zeros e uns pode omitir determinadas características das variáveis originais. Além disso, as correlações entre variáveis não são consideradas.

Aplica-se, no presente trabalho, o coeficiente de dissimilaridade baseado em distribuições condicionais de probabilidades, proposto em Quang e Bao (2005). A utilização de distribuições condicionais no cálculo do coeficiente não requer a construção de vetores binários e automaticamente incorpora associações existentes entre variáveis. A obtenção do coeficiente é realizada em duas etapas. Inicialmente, estima-se a dissimilaridade entre dois resultados k e k' de uma variável Y_i, $D_{Y_i}(k,k')$, como sendo a soma das dissimilaridades das distribuições de probabilidades das demais variáveis condicionadas nos dois resultados observados, ou seja,

$$D_{Y_i}(k,k') = \sum_{i' \neq i} \Psi\left(f\left(Y_{i'} \mid Y_i = k\right), f\left(Y_{i'} \mid Y_i = k'\right)\right),$$

sendo $f(\cdot|\cdot)$ a função de distribuição de probabilidades condicionais e $\Psi(\cdot,\cdot)$ uma medida de dessemelhança entre duas distribuições de probabilidades. Quang e Bao (2005) consideram, para tal finalidade, o método de divergência de Kullback-Leibler (Kullback e Leibler, 1951). Suponha $p(y)$ e $p'(y)$ duas funções de probabilidades quaisquer. A medida de divergência de Kullback-Leibler é calculada da seguinte forma:

$$KL(p,p') = \sum_x \left(p(y)\log_2 \frac{p(y)}{p'(y)} + p'(y)\log_2 \frac{p'(y)}{p(y)} \right),$$

Finalmente, a dissimilaridade entre dois vetores de observações $\mathbf{y}_j = (y_1, y_2, ..., y_q)$ e $\mathbf{y}_{j'} = (y_1', y_2', ..., y_q')$, denotada por $D_{jj'}$ é estimada pela soma das dissimilaridades individuais, calculadas para cada variável:

$$D_{jj'} = \sum_{i=1}^{q} D_{Y_i}\left(y_{ij}, y_{ij'}\right).$$

4 Coeficiente de entropia para múltiplas variáveis

Como discutido anteriormente, medidas de entropia podem ser utilizadas para quantificar a heterogeneidade ou impureza dos nós em árvores de classificação univariadas (Breiman *et al*., 1984). O uso da entropia em sua versão multivariada é freqüente, por exemplo, como medida de impureza dos grupos produzidos por análises de agrupamentos (Darcy e Aigner, 1980). Propõe-se aqui considerar a medida de entropia como alternativa para a construção e seleção de árvores de classificação multivariadas.

Seja $\mathbf{Y} = (Y_1, Y_2, ..., Y_q)$ um vetor de variáveis aleatórias qualitativas, cada uma com r_i possíveis resultados ($i = 1, 2, ..., q$). Desta forma, o número total de categorias da

distribuição conjunta de $\mathbf{Y}$ é $r = \prod_{i=1}^{q} r_i$. A entropia do vetor aleatório $\mathbf{Y}$ é definida da seguinte maneira:

$$H(\mathbf{Y}) = -\sum_{k=1}^{r} \left[P(\mathbf{Y} = \mathbf{y}_k) \right] \log_2 \left[P(\mathbf{Y} = \mathbf{y}_k) \right] \tag{3}$$

sendo que as probabilidades $P(\mathbf{Y} = \mathbf{y}_k)$, na prática, são estimadas pelas respectivas proporções amostrais p_k. Ressalte-se que à medida que mais variáveis respostas são consideradas, o coeficiente de entropia calculado com base na distribuição conjunta pode se mostrar inviável, dado o elevado número de resultados produzido pelas combinações das categorias de cada variável, gerando valores reduzidos para a distribuição de freqüências conjuntas.

Embora menos compatível com o contexto multivariado do estudo, a soma das entropias avaliadas individualmente para cada variável resposta também fornece um indicativo da heterogeneidade amostral, evitando o problema apontado quanto à utilização das entropias baseadas na distribuição conjunta. Nesse caso, a entropia do vetor aleatório $\mathbf{Y}=(Y_1, Y_2, ..., Y_q)$ é definida como $\sum_{i=1}^{q} H(Y_i)$, sendo $H(Y_i) = -\sum_{k=1}^{r_i} \left[P(Y_i = y_k) \right] \log_2 \left[P(Y_i = y_k) \right]$. O valor de $H(Y_i)$ está atrelado a r_i, o número de atributos da *i*-ésima variável, podendo assumir resultados no intervalo $[0, \log_2 r_i]$. A fim de evitar distorções ocasionadas pelas diferentes amplitudes dos coeficientes de entropia, padroniza-se $H(Y_i)$ da seguinte maneira:

$$H^*(Y_i) = \frac{H(Y_i)}{\log_2 r_i} = \frac{1}{\log_2 r_i} \left\{ -\sum_{k=1}^{r_i} P(Y_i = y_k) \log_2 \left[P(Y_i = y_k) \right] \right\},$$

garantindo valores de $H^*(Y_i)$ no intervalo $[0,1]$, para qualquer valor de r_i. As probabilidades $P(Y_i = y_k)$, $i = 1, 2, ..., q$; $k = 1, 2, ..., r_i$, necessárias ao cálculo de H^*, devem ser estimadas por p_{ik}, as respectivas proporções amostrais de resultados k da i-ésima variável.

5 Extensões do CART para a análise de dados categorizados multivariados com base em coeficientes de dissimilaridade e entropia

Seja $D_{jj'}$ o valor de algum coeficiente de dissimilaridade calculado a partir de dois elementos j e j', com relação a um conjunto de q atributos categorizados. Sugere-se como medida de heterogeneidade para n_t elementos que constituem um nó t a dissimilaridade média entre tais elementos, ou seja

$$\phi_{Dis}(t) = \left[\frac{n_t (n_t - 1)}{2} \right]^{-1} \sum_{j=1}^{n_t} \sum_{j'<j} D_{jj'} \tag{4}$$

A utilização da dissimilaridade média como critério de partição dos nós não caracteriza o método proposto como uma árvore de classificação segundo a definição de Breiman *et al*. (1984), uma vez que a condição de subaditividade, intrínseca à definição do CART, não é respeitada. A condição de subaditividade de uma medida de impureza $\phi(\cdot)$ garante que se um nó t for partido em dois nós t_L e t_R, então necessariamente $\phi(t) \geq \phi(t_L) + \phi(t_R)$. Isso não é válido para $\phi_{Dis}(\cdot)$. Registra-se aqui, no entanto, que outras medidas de impureza fundamentadas em dissimilaridades respeitantes à condição de subaditividade, como a soma das dissimilaridades intra-nós, foram consideradas, mas descartadas devido à produção de resultados inconsistentes.

Propõe-se a construção da árvore tomando-se como medida de impureza a dissimilaridade média, conforme definido em (4), por meio da execução das partições que proporcionem maior redução na impureza dos nós, maximizando (1). A poda é realizada com base na função de custo-complexidade apresentada em (2). A estimação do custo de má-classificação da árvore via validação cruzada requer a definição de uma medida de má-classificação, aplicada a elementos independentes dos utilizados na construção do modelo. Seja T uma árvore qualquer. Suponha que uma nova observação y^*, independente daquelas utilizadas na construção de T, seja alocada ao nó t através de T. Seja d_j^* a dissimilaridade de y^* em relação a uma observação $j \subset t$. Considera-se como medida de qualidade da predição a dissimilaridade média desta nova observação em relação às observações contidas em t, ou seja,

$$\phi_{Dis}(y^*) = \sum_{j \subset t} d_j^* / n_t \tag{5}$$

A estimação de $R(T)$ via validação cruzada é feita dividindo a amostra original (ς) em V subamostras de tamanhos (aproximadamente) iguais: $\varsigma_1, \varsigma_2, ..., \varsigma_V$. Seja $\varsigma^{(v)} = \varsigma - \varsigma_v$ a subamostra composta pelos elementos da amostra original, exceto por aqueles pertencentes a ς_v, e $T^{(v)}$ a árvore de classificação construída a partir de $\varsigma^{(v)}$, para $v = 1, 2, ..., V$. O custo de má-classificação da árvore, estimado por validação cruzada, é descrito como:

$$R_{Dis}^{CV}(T) = \sum_{v} \frac{R\left(T^{(v)}\right)}{V} \tag{6}$$

sendo $R\left(T^{(v)}\right) = \sum_{y_j \subset \varsigma_v} \phi_{Dis}(y_j) / n_v$ e n_v o número de elementos em ς_v. A seleção da melhor árvore é realizada por meio da construção do gráfico de complexidade e da aplicação da regra do desvio padrão, conforme descrito em Breiman *et al*. (1984). O desvio padrão da estimativa do custo de má-classificação é estimado com base nos resultados obtidos em dez validações cruzadas distintas.

Considerando o coeficiente de entropia para múltiplas variáveis, para um nó t qualquer, propõe-se como medidas de heterogeneidade:

$$\phi_{Ent_2}(t) = \frac{1}{q}\sum_{i=1}^{q} H_t^*(Y_i),$$

sendo $H_t^*(Y_i)$ a entropia padronizada de Y_i em t, definida em (3). A divisão por q garante valores de $\phi_{Ent_2}(t)$ no intervalo $[0,1]$.

A construção da árvore e sua poda baseiam-se, novamente, na variação da impureza causada pela partição (1) e na função de custo-complexidade (2). A seleção do modelo é executada por validação cruzada, de forma semelhante ao descrito para árvores construídas com base em coeficientes de dissimilaridades (5,6). Embora se utilize neste trabalho, com tal finalidade, o coeficiente de dissimilaridade simples, que consiste basicamente na proporção de resultados não coincidentes dentre as q variáveis sob estudo, qualquer outro coeficiente pode ser utilizado nesta etapa da análise.

A classificação dos nós finais da árvore selecionada é realizada segundo as distribuições de freqüências verificadas em cada nó. Seja $p_t(\mathbf{y})$ a distribuição de freqüências em um nó t. A classificação de um nó final t pode ser realizada segundo a distribuição conjunta, classificando-o por $\mathbf{y} = (y_1, y_2, ..., y_q)$, tal que $p_t(\mathbf{y})$ é máximo, ou segundo as distribuições marginais, classificando t por $\mathbf{y} = (y_1, y_2, ..., y_q)$ tal que $p_t(y_i)$ é máximo, $i = 1, 2, ..., q$. A classificação segundo a distribuição conjunta é inviável, novamente, devido ao elevado número de resultados produzidos pela combinação das categorias das variáveis respostas, dissipando a distribuição de freqüências e tornando instáveis as classificações resultantes. Baseado neste fato optou-se por utilizar as distribuições marginais como regra de classificação neste trabalho.

6 Delineamento do estudo por simulação aplicado à análise dos métodos multivariados de árvores de classificação

Para o estudo por simulação, foram gerados $n = 500$ vetores, compostos por três variáveis respostas multinomiais, cada uma delas com quatro categorias, e cinco covariáveis, com distribuições de probabilidades contínuas (normal e qui-quadrado) e discretas (Poison e multinomiais). A obtenção de valores amostrais para as oito variáveis foi executada, inicialmente, gerando resultados amostrais para um vetor aleatório $\mathbf{Z} = (Z_1, Z_2, ..., Z_8)$, normalmente distribuído, com vetor de médias 0 e matriz de covariâncias $\boldsymbol{\Sigma}$, sendo os elementos da diagonal de $\boldsymbol{\Sigma}$ iguais a 1. Assim, para qualquer par de variáveis, o valor da covariância equivale ao coeficiente de correlação linear de Pearson (Zar,1999). A definição de $\boldsymbol{\Sigma}$ está ligada à estrutura de dependências desejada para as variáveis a serem geradas. Considere F_i a função de probabilidades acumuladas de Z_i. Num segundo passo, calculou-se $\mathbf{U} = (U_1 = F_1(Z_1), U_2 = F_2(Z_2), ..., U_8 = F_8(Z_8))$. Dessa forma, o vetor $\mathbf{U}$ é composto por variáveis aleatórias uniformemente distribuídas

no intervalo $[0,1]$ (Ross, 1997). Para distribuições bivariadas, segundo Barnett (1980), essa forma de obtenção de distribuições uniformes conserva a dependência originalmente inserida. Finalmente, para se obter variáveis aleatórias com as distribuições de probabilidades desejadas, aplica-se às variáveis uniformes geradas, algoritmos adequados (como, por exemplo, o método da distribuição inversa – Ross, 1997). No presente estudo, quatro matrizes $\boldsymbol{\Sigma}$ foram consideradas, conforme a magnitude requerida para as covariâncias e correlações: (i) $\sigma_{ii'} = 0, \forall i \neq i'$: covariâncias (e correlações) nulas; (ii) $|\sigma_{ii'}| \leq 0{,}5, \forall i \neq i'$: covariâncias (e correlações) baixas; (iii) $|\sigma_{ii'}| \in [0,1], \forall i \neq i'$: covariâncias (e correlações) variadas e (iv) $|\sigma_{ii'}| \geq 0{,}5, \forall i \neq i'$: covariâncias (e correlações) altas.

Outra característica controlada no estudo por simulação foi a entropia das variáveis dependentes (Y_1, Y_2, Y_3). Foram consideradas variáveis com entropias baixas, geradas com vetor de probabilidades $\mathbf{p} = (0{,}895; 0{,}090; 0{,}013; 0{,}002)$, variáveis com entropias moderadas, considerando $\mathbf{p} = (0{,}623; 0{,}025; 0{,}077; 0{,}275)$ e variáveis com entropias altas, tomando $\mathbf{p} = (0{,}384; 0{,}371; 0{,}124; 0{,}121)$. Os vetores de probabilidades correspondentes a cada um dos três graus de entropia foram determinados por meio da distribuição empírica do coeficiente, estimada via simulação, e três combinações de variáveis foram consideradas: Y_1, Y_2 e Y_3 geradas com entropias baixas; Y_1, Y_2 e Y_3 geradas, respectivamente, com entropias baixa, moderada e alta e Y_1, Y_2 e Y_3 geradas com entropias altas.

Sob cada uma das 12 configurações resultantes das combinações de correlações e entropias, os dados gerados foram analisados mediante construção de modelos de classificação por árvores baseados em coeficientes de dissimilaridade e entropia. Como critérios adicionais para construção dos modelos, optou-se, com base na quantidade de elementos amostrados, por não partir nós com menos de 20 elementos e não constituir nós com menos de 10. O software estatístico R (R DEVELOPMENT CORE TEAM, 2008) foi utilizado em todas as etapas deste trabalho, desde a implementação dos algoritmos, execução do estudo por simulação e análise dos dados sobre consumo alcoólico dos habitantes do município de Botucatu.

7 Aplicação dos métodos de classificação por árvores fundamentados em coeficientes de dissimilaridades e entropia na análise de dados simulados

Os métodos propostos de classificação por árvores para dados categorizados multivariados foram avaliados quanto à entropia e à dissimilaridade média dos modelos produzidos. A Figura 1 apresenta as curvas de custo–complexidade para as entropias e dissimilaridades médias relativas aos modelos construídos com o coeficiente de dissimilaridade baseado em distribuições de probabilidades condicionais.

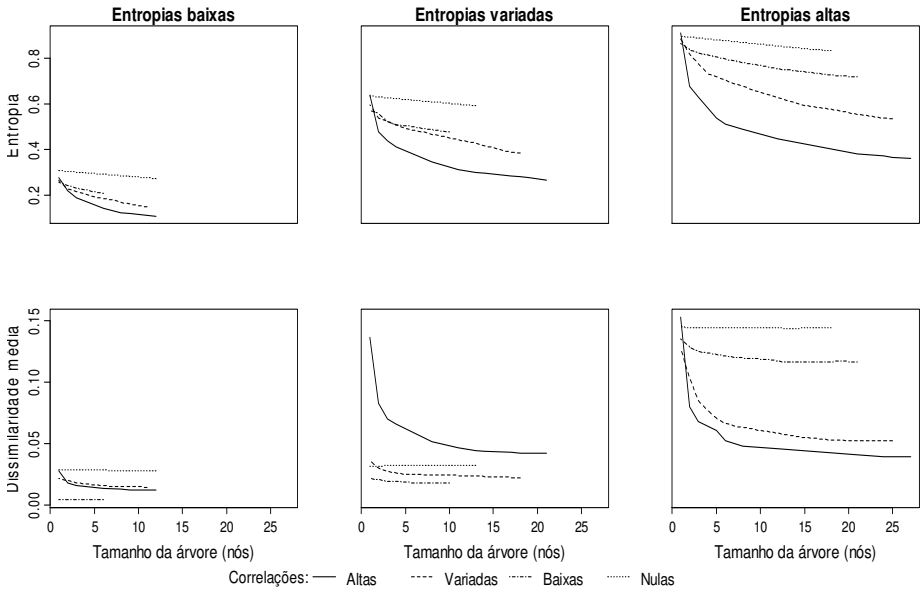

Figura 1 - Curvas de custo-complexidade para as entropias e dissimilaridades médias de modelos de classificação por árvores construídos com o coeficiente de dissimilaridade baseado em distribuições condicionais de probabilidades, para dados gerados com diferentes graus de correlação e entropia.

A influência das correlações entre variáveis respostas pode ser avaliada comparando as curvas de um mesmo gráfico, enquanto a influência das entropias de tais variáveis é avaliada comparando gráficos dispostos lado a lado em uma mesma figura. A Figura 2, por sua vez, apresenta as curvas de custo-complexidade para as 12 árvores construídas utilizando o coeficiente de entropia.

Os gráficos de custo-complexidade produzidos mediante aplicação de ambos os coeficientes evidenciam que a entropia e a dissimilaridade média proporcionada pelos modelos de classificação por árvores diminuem mais acentuadamente quanto maiores as correlações entre variáveis. Além disso, quanto maiores as entropias das variáveis respostas, maiores as reduções das medidas de heterogeneidade consideradas. Os resultados indicam, portanto, maior viabilidade da aplicação das técnicas propostas quando se dispõem de variáveis com correlações e entropias altas. Na Figura 1, a curva de custo-complexidade referente à dissimilaridade média, obtida a partir de dados gerados com entropias variadas e correlações altas, apresenta um comportamento diferenciado em relação às demais curvas, dada a maior dissimilaridade em relação às demais configurações simuladas. A construção de árvores baseadas no mesmo coeficiente de dissimilaridades, para dados gerados sob condições idênticas, não reproduziu tal comportamento.

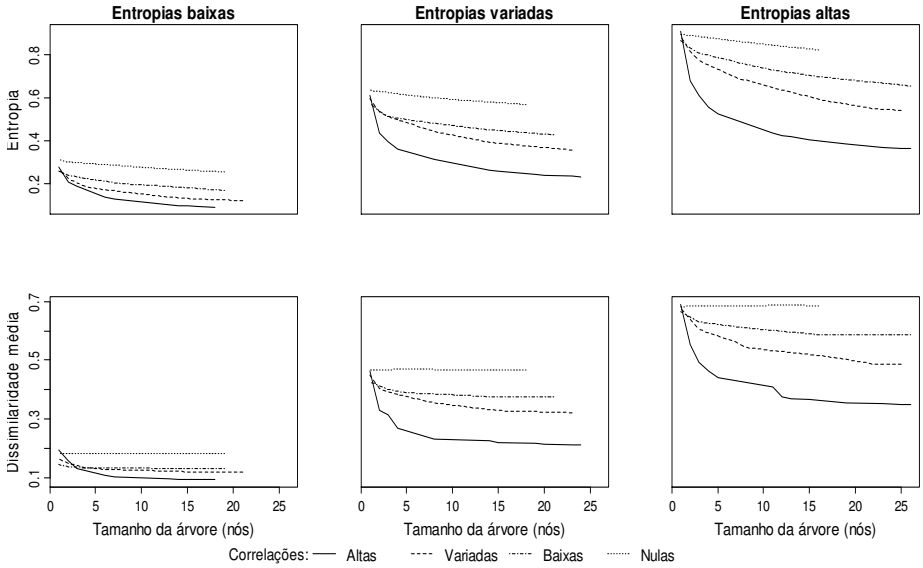

Figura 2 - Curvas de custo-complexidade para as entropias e dissimilaridades médias de modelos de classificação por árvores construídos com base no coeficiente de entropia, para dados gerados com diferentes graus de correlação e entropia.

8 Análise do perfil do consumo de álcool e fumo dentre os habitantes do município de Botucatu (SP).

As metodologias propostas de classificação multivariadas por árvores foram também aplicadas na análise de dados produzidos por um levantamento realizado no município de Botucatu (SP) como parte de estudo realizado em oito países em desenvolvimento, compondo um projeto multinacional denominado GENACIS (Gender, Alchool and Culture: an International Study). O GENACIS foi criado pela Organização Mundial de Saúde (World Health Organization – WHO), juntamente com outras agências governamentais espalhadas pelo mundo, tendo como principais objetivos a avaliação de diferenças quanto ao padrão de consumo alcoólico entre homens e mulheres, bem como a detecção de fatores pessoais, familiares e sociais associados ao consumo de álcool e as implicações do alcoolismo na saúde e no comportamento social da população. Os resultados do estudo podem ser consultados em WHO (2005).

A coleta dos dados foi realizada mediante aplicação de questionários, conduzida pelo Departamento de Saúde Pública da Universidade Estadual Paulista (UNESP), em que, no total, foram amostrados 740 indivíduos ao longo do biênio 2001-2002. No presente estudo, foram consideradas como variáveis respostas: freqüência de consumo alcoólico nos últimos 12 meses (não bebeu, poucas vezes (em menos de 12 ocasiões) ou muitas vezes (ao menos uma vez por mês)); intensidade com que consumiu álcool num único dia, quando mais bebeu nos últimos 12 meses (nada, moderada (menos de cinco drinques) ou abusiva (cinco drinques ou mais)); bebida alcoólica preferida (nenhuma, cerveja, vinho ou destilado); consumo de cigarro (sim ou não) e consumo de maconha

(sim ou não). A intensidade de consumo alcoólico é quantificada ponderando as quantidades e os tipos de bebidas citados pelo entrevistado. A Tabela 1 apresenta as 14 covariáveis consideradas, abordando características pessoais, sociais e econômicas dos entrevistados. Do total de 740 indivíduos entrevistados, 482 foram mantidos para a análise, por terem respondido a todas as perguntas do questionário.

8.1 Modelo de classificação por árvore obtido com o coeficiente de dissimilaridade baseado em distribuições condicionais de probabilidades.

A Figura 3 apresenta a curva de custo-complexidade para a seqüência de árvores aninhadas produzida mediante aplicação do coeficiente de dissimilaridades baseado em distribuições condicionais de probabilidades, indicando a seleção da árvore com 9 nós finais, de acordo com a regra do desvio padrão (Breiman *et al*, 1984). A Figura 4, por sua vez, representa a árvore selecionada. As variáveis 'ocupação profissional', 'nível educacional' e 'número de pessoas com quem o entrevistado compartilha os problemas' são responsáveis por duas partições cada, indicando a importância das mesmas na determinação do perfil dos entrevistados quanto ao consumo de álcool e fumo.

Tabela 1 - Variáveis referentes a características pessoais, componentes (ou obtidas a partir de) questionário aplicado a uma amostra de habitantes do município de Botucatu (SP) (continua)

Variável	Descrição	Tipo
GENDER	Sexo	M - Masculino F - Feminino
DATE	Ano de nascimento	Numérica
IMC	Índice de massa corporal	Numérica
SEDU	Grau máximo de escolaridade	1 – Analfabeto 2 – Alfabetizado, mas não freqüentou escola 3 – 1º grau incompleto 4 – 1º grau completo 5 – 2º grau incompleto 6 – 2º grau completo 7 – Ensino superior incompleto 8 – Ensino superior completo
SETH	Grupo étnico	1 – Branco 2 – Negro 3 – Mestiço 4 – Oriental 5 – Indígena 6 – Nenhuma das anteriores
SMST	Situação conjugal	1 – Casado 2 – Vive com parceiro 3 – Viúvo 4 – Divorciado 5 – Casado, mas separado 6 – Nunca foi casado

Tabela 1 - Variáveis referentes a características pessoais, componentes (ou obtidas a partir de) questionário aplicado a uma amostra de habitantes do município de Botucatu (SP) (continuação).

Variável	Descrição	Tipo
SNPH	Número de pessoas que residem com o entrevistado	Numérica
WPOS	Ocupação profissional atual	2 – Dona de casa 4 – Afastado por motivos de doença 5 – Aposentado 6 – Estudante 7 – Desempregado 8 – Empregado
WHHI	Renda familiar aproximada	1 – ≥ 7 salários mínimos 2 – 6 salários mínimos 3 – 5 salários mínimos 4 – 4 salários mínimos 5 – 3 salários mínimos 6 – ≤ 2 salários mínimos
NLMC	Número de contatos (e-mails, cartas, telefonemas) informais com amigos.	1 – Nenhuma vez nos últimos 30 dias 2 – 1 a 3 vezes nos últimos 30 dias 3 – 1 a 2 vezes por semana 4 – Várias vezes por semana 5 – Diariamente ou quase todos os dias
NNPI	Sem contar o parceiro conjugal, quantas pessoas têm para compartilhar seus problemas.	1 – Nenhuma 2 – Uma 3 – 2-3 4 – 4-5 5 – 6 ou mais
NRPR	Religião	1 – Nenhuma 2 – Católica 3 – Evangélica/Protestante 4 – Espírita 5 – Judeu 6 – Afro-brasileira 7 – Budista 8 – Nenhuma das anteriores
HPHH	Como o entrevistado avalia sua saúde física nos últimos 12 meses	1 – Ruim 2 – Boa
HPHH	Como o entrevistado avalia sua saúde emocional/mental nos últimos 12 meses	1 – Ruim 2 – Boa

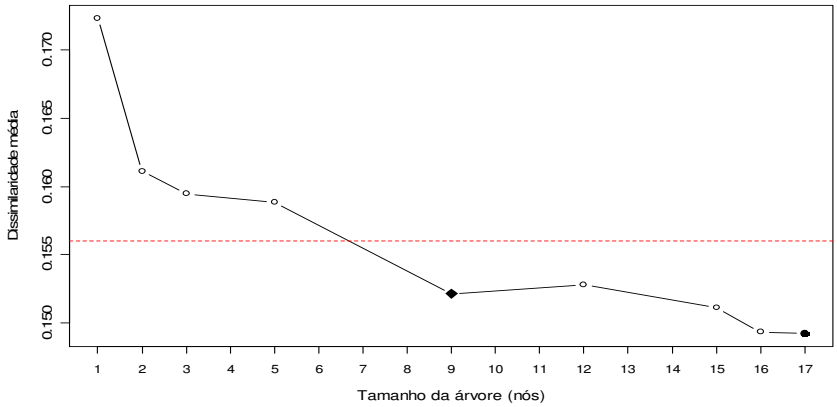

Figura 3 - Gráfico de custo-complexidade para a árvore construída para os dados de consumo de álcool e fumo, com base no coeficiente de dissimilaridade calculado a partir de distribuições condicionais de probabilidades. O ponto representado por (●) indica a árvore com menor dissimilaridade média, o ponto representado por (♦) indica a árvore selecionada pela regra do desvio padrão e a linha horizontal tracejada (---) o limite superior da dissimilaridade média associado à regra do desvio padrão.

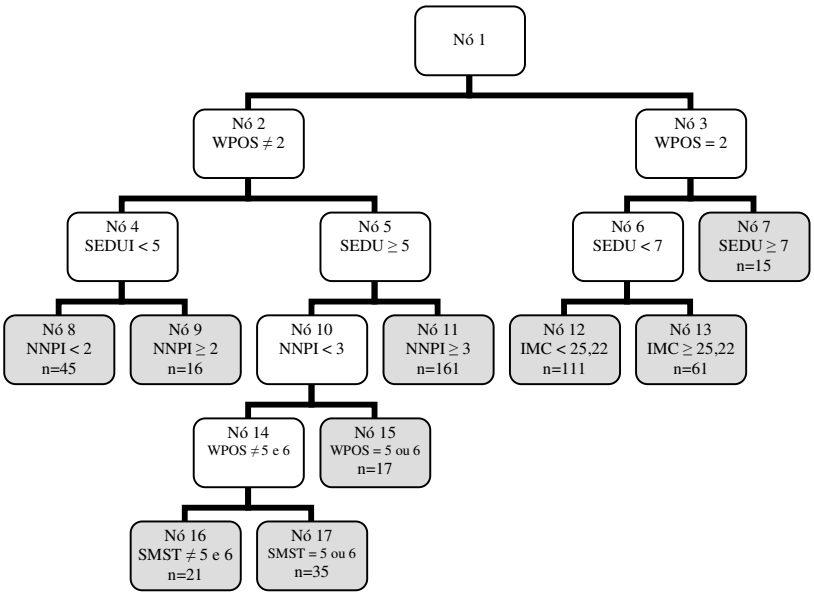

Figura 4 - Modelo de classificação por árvore obtido com o coeficiente de dissimilaridades baseado em distribuições de probabilidades condicionais. Os códigos utilizados para as variáveis que compõem o modelo são apresentados na Tabela 1. No interior de cada nó são representadas as partições executadas. Os nós com preenchimento são nós finais, apresentando, em seu interior, o número de elementos que os compõem.

A interpretação do modelo obtido requer a avaliação dos nós finais quanto às distribuições de seus componentes em relação às distribuições de freqüências de tais variáveis em cada nó final, possibilitando analisar o perfil dos nós constituídos quanto aos indivíduos que os compõem. Conclui-se, com base na árvore apresentada na Figura 4 e nos gráficos de colunas apresentados na Figura 5, que os nós 8, 11 e 15 apresentam as maiores porcentagens de indivíduos que afirmaram ter bebido muitas vezes ao longo do último ano (73%, 73% e 81%, respectivamente), e de indivíduos que afirmaram ter consumido bebidas alcoólicas de maneira abusiva (55%, 54% e 52%). O nó 8 apresenta ainda os maiores percentuais de fumantes (52%) e usuários de maconha (4%) dentre todos. Os indivíduos que compõem o nó 8 não são donas de casa, têm baixa escolaridade (no máximo completaram o primeiro grau) e afirmaram não ter com quem compartilhar seus problemas. Os indivíduos dos nós 11 e 15 têm maior escolaridade (no mínimo, segundo grau incompleto), sendo que aqueles alocados ao nó 15 são estudantes ou aposentados com no máximo uma pessoa com quem podem compartilhar seus problemas, enquanto os alocados ao nó 11 têm mais de uma pessoa para dividir as angústias.

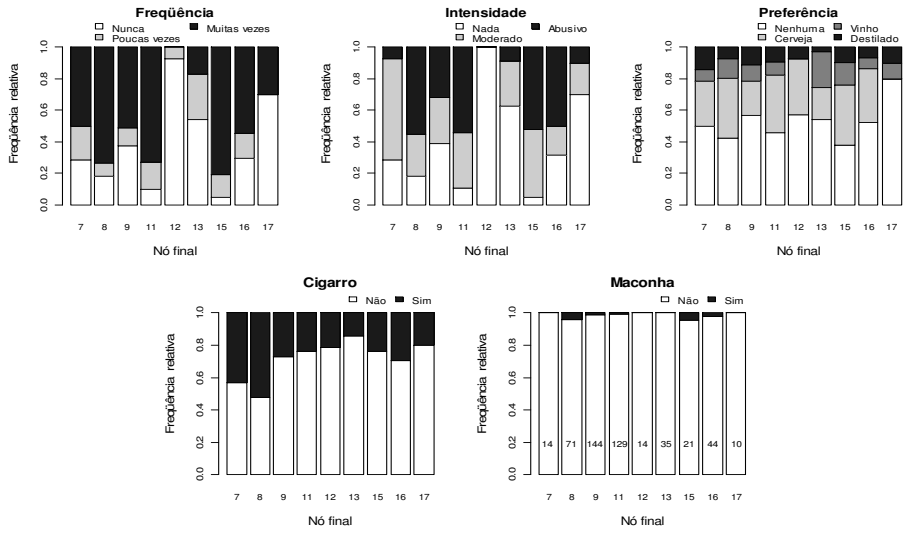

Figura 5 - Composição dos nós finais da árvore baseada no coeficiente de dissimilaridade calculado a partir de distribuições condicionais de probabilidades quanto à freqüência e à intensidade de consumo alcoólico no último ano, à bebida preferida e aos consumos de cigarro e maconha. Os valores no interior das colunas do gráfico relativo ao consumo de maconha indicam os tamanhos dos nós.

O nó 12 se destaca pela maior porcentagem de indivíduos que afirmaram não ter bebido no último ano (93%), enquanto para o nó 13 esse percentual é de 45%. Os indivíduos desses dois nós são donas de casa sem curso superior (completo ou não), diferindo, no entanto, quanto ao índice de massa corporal (inferior a 25,22 para aquelas que integram o nó 12 e superior a 25,22 para as que fazem parte do nó 13). O nó 13 apresenta o maior percentual de pessoas que têm o vinho como bebida preferida (23%).

Os indivíduos alocados aos nós 16 e 17 não são donas de casa, estudantes ou aposentados, não tem mais de uma pessoa com quem dividir os problemas e, no mínimo, deram início ao segundo grau. Diferem, no entanto, quanto à situação conjugal: aqueles que compõem o nó 17 são solteiros ou separados, enquanto os que compõem o nó 16 não são nem solteiros nem separados. Comparando os dois nós, tem-se que indivíduos do nó 17 bebem com mais freqüência (54% afirmaram ter bebido muitas vezes ao longo do último ano, contra 30% do nó 16), fumam mais (30% de fumantes, contra 20% do nó 17), e bebem com mais intensidade (50% afirmaram ter abusado ao menos uma vez, contra 10% dos indivíduos do nó 17).

A Figura 6 apresenta o gráfico produzido por uma análise de correspondência múltipla (GREENACRE, 2007) compreendendo as cinco variáveis dependentes e uma variável indicadora dos nós aos quais os indivíduos foram alocados. Proximidades entre as representações de categorias de diferentes variáveis, no gráfico de análise de correspondência, indicam maior freqüência observada para a combinação de tais categorias do que o esperado na situação de independência. A proximidade dos nós 12 e 13 às categorias associadas ao não consumo de álcool, a representação do nó 8 no mesmo quadrante do consumo freqüente e abusivo de álcool e fumo, e a maior proximidade do nó 16 às categorias de consumo alcoólico do que o nó 17 confirmam as evidências levantadas anteriormente.

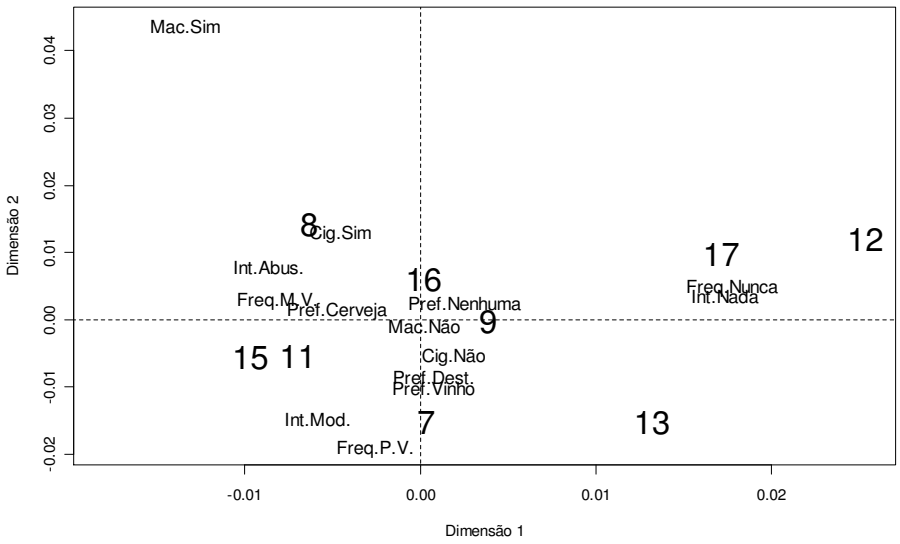

Figura 6 - Gráfico da análise de correspondência múltipla para as variáveis 'Freqüência com que bebeu no último ano' – Freq. (sendo M.V. = muitas vezes e P.V. = poucas vezes), 'Intensidade com que bebeu quando mais consumiu álcool' – Int. (sendo Mod. = Moderado e Abus. = abusivamente), 'Consumo de cigarro' – Cig., 'Consumo de maconha' – Mac. e 'Bebida preferida' – Pref. (sendo Dest. = destilado). Os números representados no interior do gráfico indicam os nós finais.

8.2 Árvore de classificação multivariada obtida com o coeficiente multivariado de entropia.

A curva de custo-complexidade produzida pela seqüência de árvores aninhadas obtidas com a aplicação do coeficiente de entropia encontra-se na Figura 7, indicando, segundo a regra do desvio padrão, a seleção da árvore com 11 nós finais. A árvore selecionada é representada na Figura 8. A Figura 9, por sua vez, apresenta os resultados da análise de correspondência múltipla, aplicada ao conjunto de variáveis respostas, acrescido da variável indicadora dos nós finais aos quais as observações foram alocadas.

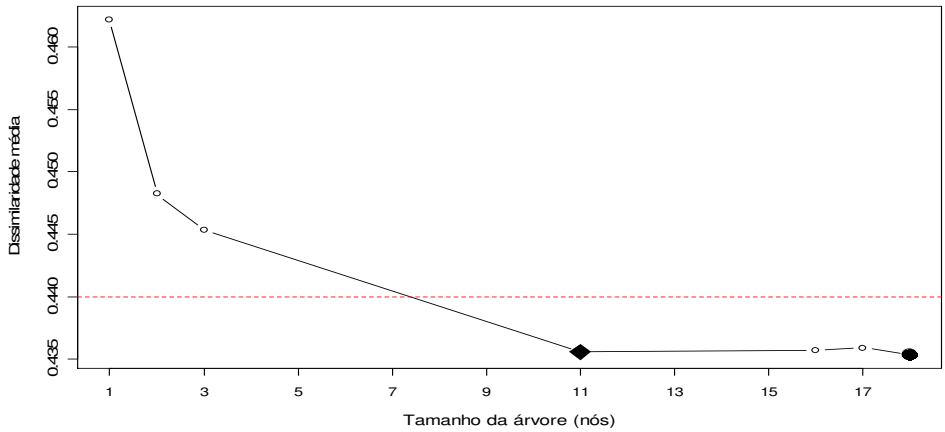

Figura 7 - Gráfico de custo-complexidade para a árvore de classificação multivariada construída para os dados de consumo de álcool e fumo, com base na medida de entropia. O ponto representado por (●) indica a árvore com menor dissimilaridade média, o ponto representado por (♦) indica a árvore selecionada pela regra do desvio padrão e a linha horizontal tracejada (---) o limite superior da dissimilaridade média associado à regra do desvio padrão.

Uma primeira conclusão extraída do gráfico da análise de correspondência apresentado na Figura 9 refere-se ao fato de o consumo de cigarro, maconha e consumo freqüente e abusivo de álcool terem suas representações no mesmo, indicando associação entre tais categorias. Além disso, os nós 18, 19 e 20 também estão representados neste quadrante, o que evidencia associação entre estes três nós e as categorias mencionadas. O nó 18 é composto por indivíduos que não são aposentados ou donas de casa, não têm com quem compartilhar os problemas, não têm curso superior, residem com mais de quatro pessoas e têm IMC inferior a 21,6. Os indivíduos que compõem o nó 19 têm perfil idêntico, mas IMC superior a 21,6. Já os indivíduos que compõem o nó 20 também não são donas de casa ou aposentados, têm duas pessoas ou mais com quem dividir os problemas, são viúvos, divorciados ou separados e tem IMC inferior a 22,98.

As categorias referentes ao consumo alcoólico moderado e pouco freqüente, além da preferência por cerveja, têm suas representações num mesmo quadrante, juntamente com

os nós 7, 21 e 16, indicando que os elementos que compõem os referidos nós bebem poucas vezes e com moderação. Os indivíduos que compõem o nó 7 são donas de casa e aposentados com curso superior. Já aqueles que compõem o nó 21 têm características semelhantes às mencionadas para o nó 20, mas com IMC superior a 22,98. Quanto ao nó 16, pode-se caracterizar seus componentes por não serem aposentados ou donas de casa, terem duas pessoas ou mais com quem dividir seus problemas e serem casados, viverem com parceiro ou nunca terem se casado.

O nó 13 está associado ao não consumo de bebidas alcoólicas, o que pode ser verificado pela proximidade de sua representação, no gráfico de análise de correspondência, em relação às categorias relativas ao não consumo de bebida alcoólica. De forma um pouco menos acentuada indivíduos do nó 12 também são avessos ao consumo de álcool, cigarro e maconha. O nó 13 é composto por donas de casa sem curso superior, enquanto o nó 12 é formado por aposentados sem curso superior. Os gráficos de distribuição de freqüências que compõem a Figura 10 apresentam as composições de cada nó quanto às variáveis de consumo alcoólico e fumo e dão suporte para as conclusões citadas anteriormente, baseadas nos resultados da análise de correspondência.

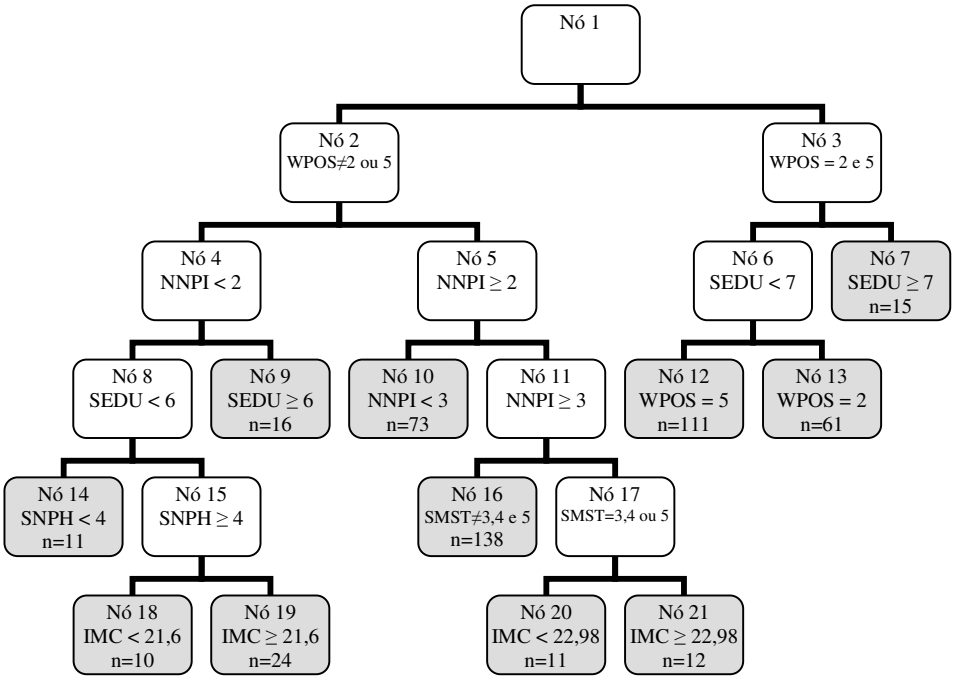

Figura 8 - Árvore de classificação multivariada obtida com o coeficiente de entropia. Os códigos utilizados para as variáveis que compõem o modelo são apresentados na Tabela 1. No interior de cada nó são representadas as partições executadas. Os nós com preenchimento são nós finais, apresentando, em seu interior, o número de elementos que os compõem.

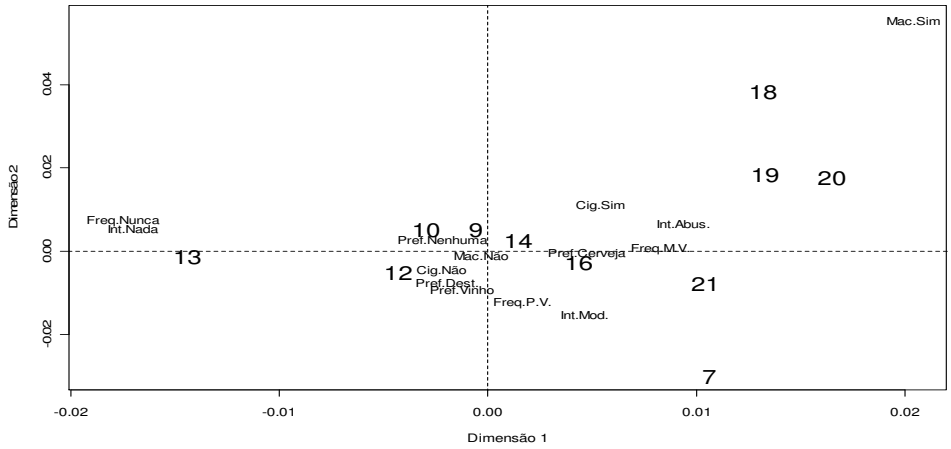

Figura 9 - Gráfico da análise de correspondência múltipla para as variáveis 'Freqüência com que bebeu no último ano' – Freq (sendo MV = muitas vezes e PV = poucas vezes), 'Intensidade com que bebeu quando mais consumiu álcool' – Int (sendo Mod = Moderado e Abus = abusivamente), 'Consumo de cigarro' – Cig, 'Consumo de maconha' – Mac e 'Bebida preferida' – Pref (sendo Dest = destilado). Os números representados no interior do gráfico indicam os nós finais.

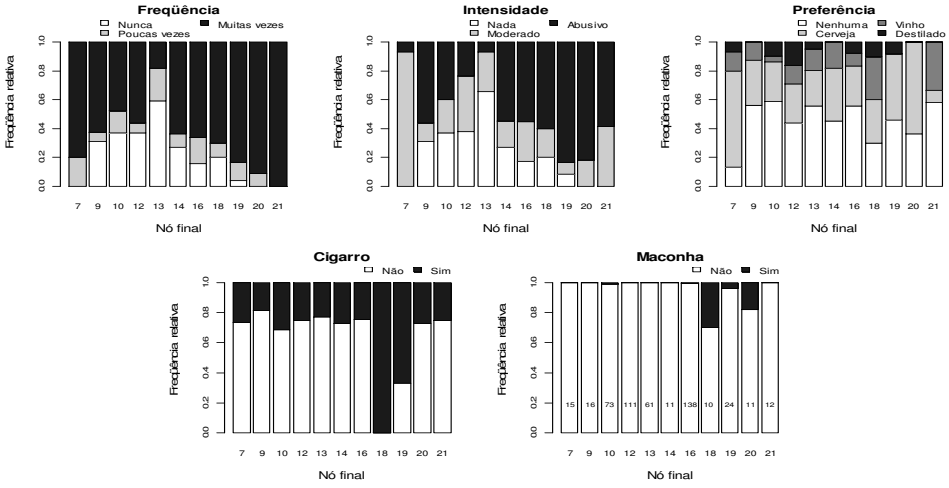

Figura 10 - Composição dos nós finais da árvore de classificação multivariada baseada na medida de entropia, quanto à freqüência e à intensidade de consumo alcoólico no último ano, à bebida preferida e aos consumos de cigarro e maconha. Os valores no interior das colunas do gráfico relativo ao consumo de maconha indicam os tamanhos dos nós.

Conclusões

Pretendeu-se, por meio deste trabalho, conceber novas técnicas de modelagem adequadas à análise de dados multivariados categorizados, por meio da proposição de procedimentos multivariados baseados em árvores de classificação. Tais procedimentos, fundamentados em coeficientes de dissimilaridade e entropia, foram apresentados e tiveram seus desempenhos avaliados com base em um estudo por simulação e em suas aplicações na análise de dados de consumo alcoólico e fumo dentre habitantes do município de Botucatu (SP).

Pôde-se verificar, por meio dos resultados apresentados, que os métodos propostos são capazes de explicar a variação original dos dados, sobretudo quando são analisadas variáveis com correlações e entropias moderadas ou altas. A análise dos dados de consumo de álcool e fumo permitiu detectar perfis diferentes de indivíduos que se associam a padrões distintos de consumo de álcool e fumo. Para os dois modelos construídos, um baseado em coeficiente de dissimilaridade e outro de entropia, variáveis como 'ocupação profissional atual', 'grau máximo de escolaridade' e 'número de pessoas com quem pode compartilhar os problemas' mostram-se importantes na explicação do conjunto de variáveis relacionadas aos consumos de álcool e fumo. Tais covariáveis são responsáveis, cada uma delas, por duas partições em ambos os modelos. As conclusões extraídas dos dois modelos são compatíveis.

Agradecimentos

À Prof.ª Florence Kerr-Correa por disponibilizar os dados para análise; ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro em forma de bolsa de estudos.

TACONELI, C. A.; ZOCCHI, S. S.; DIAS, C. T. S. Extensions of the algorithm of classification trees for data analysis categorized using multivariate coefficients of dissimilarity and entropy. *Rev. Bras. Biom.*, São Paulo, v.27, n.1, p.91-114, 2009.

- *ABSTRACT: The statistical analysis of large datasets requires the use of flexible methodologies, that can provide insight and understanding even in the presence of difficulties such as large numbers of variables having variable levels of association between themselves, and missing data. The construction of classification and regression trees allows for modeling of a categorical or numerical response variable as a function a set of covariates, while bypassing many of the cited difficulties. Multivariate trees extend classification and regression techniques to allow for joint analysis of two or more response variables. In recent studies, application of multivariate classification and regression techniques has been most common in situations involving numerical response variables. In this work we propose alternatives for constructing multivariate classification trees for multiple categorized response variables. Such alternatives are based on dissimilarity and entropy measures. A simulation study was used to examine the effect of variable correlations and entropies on the performance of the proposed methodology (results are better for high correlations and entropies). Analysis of data on alcohol consumption and smoking among inhabitants from Botucatu (SP) complements the analysis by showing that factors as the education level, daily occupation and possibility of sharing problems with friends have an influence on the alcohol consumption and smoking.*

▪ *KEYWORDS: Classification trees; Dissimilarity; Entropy; Alcohol and smoking; Multivariate simulation*

Referências

BARNETT, V. Some bivariate uniform distributions. *Commun. Stat. – Part A Theory Methods*, New York, v.9, n.4, 453-461, 1980.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. *Classification and regression trees*. California: Wadsworth International Group, 1984. 358p.

COX, F.; COX, A. A. *Multidimensional scaling*. 2. ed. Boca Raton: Chapman & Hall, 2001. 318p.

DARCY, R. AIGNER, H. The uses of entropy in the multivariate analysis of categorical variables. *Am. J. Polit. Sci.*, Austin, v.24, n.1, p.155-174, 1980.

DE'ATH, G. Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology,* Brooklin, v.83, n.4, p.1105–1117, 2002.

DE'ATH, G.; FABRICIUS, K. E. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, Brooklin, v.81, n.11, p.3178–3192, 2000.

GREENACRE, M. J. *Correspondence Analysis in pratice*. 2. ed. London: Academic Press, 2007. 280p.

KULLBACK, S.; LEIBLER, R.A. On Information and Sufficiency. A*nn. Math. Stat.*, Beachwood, v.22, n.1, p.79-86, 1951.

LARSEN D. R., SPECKMAN P. L. Multivariate regression trees for analysis of abundance data. *Biometrics*, Washington, v.60, n.2, p.543–549, 2004.

LEE, S. K. On generalized multivariate decision tree by using GEE. *Comput. Stat. & Data Anal.*, Amsterdan, v.49, n.4, p.1105-1119, 2005.

MILLER, J.; FRANKLIN, J. Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. *Ecol. Modell.* Amsterdan, v.157, n.2-3, p.227-247, 2002.

QUANG, L. S.; BAO, H. T. An association-based dissimilarity measure for categorical data. *Pattern Recog. Lett.***,** Amsterdam, v.26, p.2549-2557, 2005.

R DEVELOPMENT CORE TEAM. R: A LANGUAGE AND ENVIRONMENT FOR STATISTICAL COMPUTING, Vienna, Austria, 2008. Disponível em: http://www.R-project.org, 2008.

ROSS, S. M. *Introduction to probability models*. 6. ed. San Diego: Academic Press, 1997. 669p.

SEGAL, M.R. Tree-structured methods for longitudinal data. *J. Am. Stat. Assoc.*, Boston, v.87, p.407–418, 1992.

WHO. *Alcohol, gender and drinking problems: perspectives from low and middle income countries*. Genebra: Isidore S. Obot & Robin Room, 2005, 227p.

ZAR, J. H. *Biostatistical analysis.* 4. ed. New Jersey: Prentice Hall, 1999. 663p.

ZHANG, H.P. Classification trees for multiple binary responses, *J. Am. Stat. Assoc.*, Boston, v.93, p.180-193, 1998.

Recebido em 17.02.2009.

Aprovado após revisão 30.05.2009.