Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Árvores de classificação multivariadas fundamentadas em coeficientes de dissimilaridade e entropia

Cesar Augusto Taconeli

Tese apresentada para obtenção de título de Doutor em Agronomia. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2008

Cesar Augusto Taconeli
Bacharel em Estatística

Árvores de classificação multivariadas fundamentadas em coeficientes de dissimilaridade e entropia

Orientador:
Prof. Dr. **Silvio Sandoval Zocchi**

Tese apresentada para obtenção de título de Doutor em Agronomia. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2008

Dedicatória

A Deus, pelas bênçãos e graças diárias,

aos meus pais, Antonio e Alzira, e aos meus irmãos,
Fábio e João Paulo, pelo apoio e incentivo incondicionais.

AGRADECIMENTOS

Ao meu orientador, Prof. Silvio Sandoval Zocchi, pela colaboração dispensada, pela motivação constante e pela confiança em meu trabalho.

Ao Prof. Carlos Tadeu dos Santos Dias, pela disponibilidade em nos ajudar nos momentos de dúvida.

Aos professores e funcionários do departamento de Ciências Exatas da ESALQ/USP e do departamento de Bioestatística do IBB/UNESP, pela amizade, incentivo e presteza em nos auxiliar sempre que necessário.

À Prof.ª Lúcia Pereira Barroso, ao Prof. Rodolfo Hoffman e à Prof.ª Clarice Garcia Borges Demétrio, pelas valiosas contribuições em meu exame de qualificação.

À Prof.ª Luzia Aparecida Trinca e ao Prof. Carlos Roberto Padovani, pelo empenho na procura e seleção de dados para ilustrar minha pesquisa.

À Prof.ª Florence Kerr-Correa por disponibilizar os dados para análise.

Ao Prof. José Cláudio Faria, pelas importantes sugestões em programação.

Ao Prof. Marco Aurélio Rodriguez Gastonguay e ao Prof. Celestin C. Kokonedji, pelas revisões em traduções para a língua inglesa.

A Osmar Jesus Macedo, a David José Miquelluti, a Lúcio Borges de Araújo e a Milton Yoshio Saito, pelas parcerias e convivências amigáveis em república.

Aos amigos de turma, Afrânio Márcio Corrêa Vieira, Andréia da Silva Meyer, Giovana Oliveira Silva e Júlio Cesar Pereira, companheiros que tornaram esta jornada muito mais agradável.

Aos colegas do programa de Pós-Graduação em Estatística e Experimentação Agronômica, pelos bons momentos em suas companhias.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro em forma de bolsa de estudos.

A todos que direta ou indiretamente contribuíram com a execução deste trabalho.

SUMÁRIO

RESUMO 7
ABSTRACT 8
1 INTRODUÇÃO 9
2 REVISÃO DA LITERATURA 11
2.1 Árvores de classificação e regressão 11
2.2 Árvores de classificação e regressão multivariadas 15
2.3 Coeficientes de similaridade e dissimilaridade para múltiplas variáveis categorizadas 17
2.3.1 Coeficientes de similaridade e dissimilaridade para múltiplas variáveis binárias 17
2.3.2 Coeficiente de dissimilaridade simples para múltiplas variáveis categorizadas 19
2.3.3 Coeficiente de dissimilaridade baseado em probabilidades 19
2.3.4 Coeficiente de dissimilaridade baseado em distribuições condicionais de probabilidades 22
2.4 Medida de entropia para múltiplas variáveis 23
2.5 Geração de valores amostrais para a distribuição multinomial e para distribuições multivariadas de probabilidades 24
2.5.1 Geração de valores amostrais para uma variável aleatória com distribuição multinomial 24
2.5.2 Geração de valores amostrais para variáveis aleatórias multivariadas correlacionadas 25
2.6 Análise de correspondência 27
3 METODOLOGIA 28
3.1 Material 28
3.2 Métodos 31
3.2.1 Geração de valores amostrais para variáveis multinomiais dependentes 31
3.2.2 Construção de árvores de classificação multivariadas com base em coeficientes de dissimilaridade 33
3.2.3 Construção de árvores de classificação multivariadas com base em medida de entropia 36

3.2.4 Seleção do modelo de árvore de classificação pelo ponto da curva de custo complexidade mais afastado de uma reta representandovariação uniforme do custo 38
3.2.5 Delineamento do estudo por simulação aplicado à análise dos métodos multivariados de árvores de classificação 38
4 RESULTADOS 40
4.1 Geração de valores amostrais para múltiplas variáveis aleatórias multinomiais..... 40
4.2 Árvores de classificação multivariadas baseadas em coeficientes de dissimilaridade e entropia para dados gerados de variáveis com diferentes graus de correlação e entropia..... 43
4.2.1 Comparação dos procedimentos uni e multivariados na classificação dos nós finais em árvores multivariadas..... 48
4.2.2 Avaliação do critério do ponto mais afastado na seleção de árvores de classificação multivariadas 49
4.3 Árvores de classificação multivariadas aplicadas ao estudo do consumo de álcool e fumo dentre os habitantes do município de Botucatu (SP). 50
4.3.1 Árvore de classificação multivariada para os dados de consumo de álcool e fumo, obtida com os coeficientes de dissimilaridade simples e baseado em probabilidades. 51
4.3.2 Árvore de classificação multivariada para os dados de consumo de álcool, cigarro e maconha fundamentada no coeficiente de dissimilaridade baseado em distribuições condicionais de probabilidades..... 58
4.3.3 Árvore de classificação multivariada para os dados de consumo de álcool, cigarro e maconha fundamentada no coeficiente de entropia..... 63
5 CONSIDERAÇÕES FINAIS 70
REFERÊNCIAS 72
APÊNDICE 75

RESUMO

Árvores de classificação multivariadas fundamentadas em coeficientes de dissimilaridade e entropia

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 construção 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 visa permitir a análise conjunta de duas ou mais variáveis respostas. Embora seja objeto de estudos recentes, a proposição de técnicas multivariadas de classificação e regressão por árvores tem sido verificada de maneira mais acentuada para situações em que se dispõe de múltiplas variáveis respostas numéricas. Propõem-se, neste trabalho, novas alternativas para a construção de árvores de classificação multivariadas, visando analisar múltiplas variáveis respostas categorizadas. Tais alternativas baseiam-se em medidas de dissimilaridade e entropia. Por meio de um estudo de simulação, verificou-se o efeito das correlações e entropias das variáveis no desempenho das metodologias propostas (os resultados 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, evidenciando, dentre outras coisas, que fatores como o grau de escolaridade, a ocupação profissional e a possibilidade de compartilhar problemas com amigos têm influência sobre os consumos de álcool e fumo dos habitantes.

Palavras chave: Árvores de classificação; Dissimilaridade; Entropia; Álcool e fumo; Simulação multivariada

ABSTRACT

Multivariate classification trees based on dissimilarity and entropy coefficients

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

1 INTRODUÇÃO

Em experimentos e levantamentos, muitas vezes, são produzidos dados complexos, com grandes números 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 proposição de métodos exploratórios capazes de produzir resultados de fácil compreensão em tais ocasiões torna-se, então, fundamental. Neste contexto os modelos de classificação e regressão por árvores ("Classification And Regression Trees" – CART - BREIMAN et al., 1984; DE'ATH e FABRICIUS, 2000) aparecem como uma alternativa exploratório/preditiva de grande valia, devido à simplicidade e versatilidade de tais técnicas.

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 sucessivas partições binárias de uma amostra, com base nos resultados 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 outros 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, análise de agrupamentos, dentre outros métodos estatísticos. Além disso, o CART destaca-se por sua flexibilidade, não impondo 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 respostas consiste em levar em consideração possíveis correlações entre as mesmas mediante construção de um único modelo. A possibilidade de analisar dados multivariados, com múltiplas covariáveis, sem qualquer restrição paramétrica, constitui forte atrativo do CART multivariado.

A construção de árvores de classificação e regressão multivariadas requer, no entanto, critérios adequados quanto à segmentação das amostras e avaliação da qualidade do modelo.

O presente trabalho tem como objetivo propor técnicas para a construção de árvores de classificação multivariadas, baseadas em coeficientes de dissimilaridade e entropia. Tais técnicas visam facilitar o estudo de dados categorizados multivariados explicados por um conjunto de covariáveis, sejam categorizadas ou numéricas. Os procedimentos propostos são avaliados por meio de um estudo de simulação, e uma aplicação com dados de consumo alcoólico, cigarro e maconha, produzidos mediante aplicação de um questionário a uma amostra de habitantes do município de Botucatu (SP) complementa a análise.

2 REVISÃO DA LITERATURA

Neste capítulo, são apresentados os principais aspectos relativos à construção de árvores de regressão e classificação univariadas (Seção 2.1) e multivariadas (Seção 2.2). Na seqüência, a Seção 2.3 apresenta coeficientes de dissimilaridade para dados categorizados, enquanto a Seção 2.4 se atém à exposição de medida de entropia para múltiplas variáveis. A Seção 2.5, por sua vez, apresenta técnicas de geração de valores amostrais para a distribuição multinomial e para distribuições multivariadas. Finalizando, a Seção 2.6 descreve resumidamente a técnica de análise de correspondência múltipla, aqui aplicada com finalidade exploratória para os resultados produzidos por árvores de classificação multivariadas.

2.1 Árvores de classificação e regressão

Certos termos são utilizados de forma recorrente na caracterização dos componentes de uma árvore de regressão ou classificação. Denomina-se *nó inicial* à amostra original, *nós intermediários* às subamostras que originam novas subamostras e *nós finais* às subamostras não partidas. Denota-se por t um nó qualquer. Além disso, as partições executadas podem ser denominadas *ramos*. A árvore é a representação gráfica de nós e ramos. Tais termos são representados na Figura 1, que ilustra uma árvore de classificação/regressão.

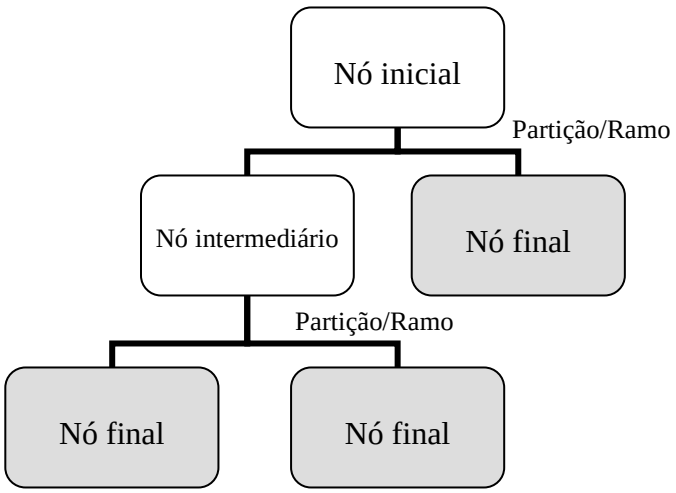

Figura 1 – Ilustração de uma árvore de classificação/regressão.

A construção de uma árvore de classificação ou regressão é dividida basicamente em quatro etapas: definição (e execução) de um critério de partição das amostras, aplicação do processo de poda, seleção do melhor modelo e classificação dos nós finais. Tais etapas são detalhadas na seqüência, inicialmente no contexto univariado.

Como dito anteriormente, a árvore é o resultado final de partições binárias sucessivas de uma amostra original. Tais partições devem ser executadas com base nos resultados verificados para as covariáveis, buscando reduzir ao máximo a variação dos resultados nas subamostras constituídas em relação à variação avaliada no nó original. Para tanto, devem-se estabelecer critérios para comparação e seleção de partições amostrais.

A partição de uma amostra se dá de maneira diferente para covariáveis ordenáveis (quantitativas ou qualitativas ordinais) ou não ordenáveis. Seja $\left\{ Y_j, \mathbf{X}_j \right\}, j = 1,2,...,n$, uma amostra de tamanho n de uma variável dependente Y e de um vetor *p*-dimensional de covariáveis $\mathbf{X} = \left(X_{1j} , X_{2j} , ... , X_{pj} \right)$. Considere X_l uma variável ordenável e τ um dos resultados amostrados para X_l. Parte-se a amostra em duas alocando elementos a nós distintos conforme resposta positiva ou negativa à questão $"X_{lj} \leq \tau ?"$. Repete-se o procedimento para os demais resultados amostrados para X_l. Caso a variável X_l não seja ordenável, o seguinte procedimento deve ser adotado: seja A uma categoria (ou subconjunto de categorias) de X_l. Alocam-se elementos a nós distintos conforme resposta (positiva ou negativa) à questão $"X_{lj} \in A?"$. De maneira similar, repete-se o procedimento para as demais categorias (e subconjuntos de categorias) amostradas para X_l.

Algumas restrições, no entanto, devem ser consideradas quanto à partição de um nó. É adequado, por exemplo, determinar o número mínimo de elementos para que um nó possa ser partido. Subdividir nós com reduzido número de elementos pode aumentar a complexidade do modelo e afetar sua capacidade preditiva, dada a possível instabilidade dos nós formados. Problemas semelhantes podem ocorrer caso nós com poucos elementos sejam originados. Por isso, é também recomendável limitar o número mínimo de elementos nos nós formados.

Respeitadas tais restrições, tem-se, para as *p* covariáveis, um conjunto de possíveis partições, que devem ser comparadas, executando-se aquela responsável por uma maior redução da heterogeneidade dos nós. Para isso, precisa-se definir alguma medida de heterogeneidade, a

fim de quantificar a variação dos dados dentro de um nó t qualquer. Breiman et al. (1984) definem e apresentam, com tal finalidade, diferentes coeficientes denominados *medidas de impureza*. Para árvores de classificação, apresentam a medida de entropia, referenciada em Zar (1999) como medida de diversidade de Shannon. Seja Y uma variável categorizada e $\{Y_1, Y_2, \ldots, Y_r\}$ seu conjunto de resultados possíveis. Define-se a medida de entropia em um nó t como

$$\phi(t) = -\sum_{k=1}^{r} p_t(y_k) \log_2\left(p_t(y_k)\right),$$

sendo $p_t(y_k)$ a proporção de elementos alocadas ao nó *t* pertencentes à classe *k*. Caso $p_t(y_k) = 0$, considera-se $p_t(y_k) \log_2\left(p_t(y_k)\right) = 0$. Nesse caso, tem-se que $Mín(\phi(t)) = 0$, quando $p_t(y_k) = 1$ e $p_t(y_{k'}) = 0$, $\forall\, k' \neq k$. Além disso, tem-se que $Máx(\phi(t)) = \log_2 k$, quando $p_t(y_k) = 1/k$, $\forall\, k \in \{1, 2, \ldots, r\}$.

Para árvores de regressão, sugerem a utilização do índice de análise de variância (índice ANOVA). Seja Y uma variável numérica. O índice ANOVA é a soma de quadrados dos desvios dos elementos presentes no nó t em torno de sua média, ou seja,

$$\phi(t) = \sum_{j} \left(y(j \mid t) - \bar{y}(t)\right)^2,$$

sendo $y(j \mid t)$ o resultado associado à observação j contida em t e $\bar{y}(t)$ a média dos elementos em t.

Suponha, agora, que o nó t seja dividido em dois novos nós (t_L e t_R), segundo uma partição s. A variação de heterogeneidade ocasionada por s é definida como

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

Deve-se, portanto, executar a partição s responsável por maximizar $\Delta_\phi(s,t)$. As subamostras originadas devem ser sucessivamente partidas, com base no critério de partição estabelecido, até a constituição de uma árvore com reduzido número de elementos em cada nó final.

Em árvores de regressão e classificação, assim como em qualquer outro procedimento de modelagem estatística, deve-se buscar um modelo parcimonioso, no caso, uma árvore de tamanho (número de nós finais) reduzido, com baixa heterogeneidade em seus nós finais e elevada capacidade preditiva. Construída a árvore, deve-se iniciar o procedimento de poda, que consiste em desfazer, uma a uma, aquelas partições que menos contribuem para a explicação da variável resposta.

O processo de poda parte da definição de uma função do tipo custo-complexidade (BREIMAN et al., 1984). Sejam T_{MAX} a maior árvore construída inicialmente e $\tilde{T}$ o conjunto de nós finais para uma subárvore T qualquer de T_{MAX}. Sejam, ainda, $\left|\tilde{T}\right|$ o número de nós finais de T e $\alpha \geq 0$ uma constante real denominada *parâmetro de complexidade*. Define-se uma medida de *custo complexidade* como

$$R_\alpha(T) = R(T) + \alpha\left|\tilde{T}\right|,$$

sendo $R(T) = \sum_{t \in \tilde{T}} \phi(t)$ o custo associado à taxa de má-classificação da árvore T e $\phi(t)$ alguma medida de heterogeneidade calculada em um nó $t \in T$. Logo, $R_\alpha(T)$ é uma combinação linear de $R(T)$ e $\left|\tilde{T}\right|$, ou seja, pondera o custo de má-classificação e o número de nós finais da árvore. 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 poda é finalizada com a obtenção da seqüência completa, iniciada pela árvore originalmente formada até a conjunção de todos os elementos em um único nó.

A comparação das árvores, dentro da seqüência aninhada, pode ser realizada por meio dos custos de má-classificação das mesmas. Fixado α, a função custo-complexidade é minimizada pela árvore $T(\alpha)$ que satisfaz às condições:

i. $R_\alpha[T(\alpha)] = \underset{T \leq T_{MÁX}}{Mín}\, R_\alpha(T)$;

ii. Se $R_\alpha(T) = R_\alpha[T(\alpha)]$, então $T(\alpha) \leq T$,

sendo $R_\alpha[T(\alpha)]$ o custo de má classificação da árvore, estimado por validação cruzada. A segunda condição favorece a seleção da árvore de menor tamanho responsável pela minimização da função de custo-complexidade. Detalhes do procedimento de validação cruzada serão apresentados posteriormente, já inseridos no contexto de árvores de classificação multivariadas.

A produção de um gráfico de $R(T)$ versus $|T|$ permite avaliar a variação do custo de má-classificação da árvore conforme aumenta sua complexidade, e a comparação dos resultados serve como subsídio para a determinação do melhor modelo. Breiman et al. (1984) propõem a seleção da menor árvore responsável por uma taxa de erro estimada que esteja a menos de um desvio padrão, relativo ao estimador do custo adotado, da menor taxa de erro verificada.

Escolhida uma árvore, a caracterização dos nós finais é realizada pela classe que aparece com maior freqüência dentre os elementos que constituem o nó, caso a variável dependente seja categorizada, ou pela média dos elementos que constituem o nó, caso seja numérica. Caracterizados os nós finais, a predição de novos elementos pode ser realizada conduzindo-os pela árvore e inferindo suas respostas de acordo com o resultado característico do nó final ao qual foram alocados.

2.2 Árvores de classificação e regressão multivariadas

Na seção anterior, considerou-se a construção de modelos de regressão e classificação univariados, ou seja, com apenas uma variável resposta. Em muitos casos, no entanto, tem-se por objetivo avaliar o comportamento de $q > 1$ variáveis com base nos resultados de um conjunto de covariáveis. A análise e a interpretação conjunta de modelos univariados para cada variável resposta é inviável à medida que o número de análises aumenta e, sobretudo, possíveis correlações entre tais variáveis são ignoradas. A proposição e a aplicação de metodologias multivariadas tornam-se, portanto, fundamentais.

Pioneiro na extensão do CART para respostas múltiplas, Segal (1992) propõe procedimentos de regressão por árvores para dados longitudinais. Tais procedimentos baseiam-se na segmentação dos elementos a partir de sua estrutura de médias, utilizando, por exemplo, a estatística T^2 de Hotteling (JOHNSON E WICHERN, 1998), ou de sua estrutura de covariâncias (nesse caso, o autor sugere quantificar a heterogeneidade dos dados a partir dos resíduos obtidos após extrair o efeito das médias).

A construção de árvores de classificação para respostas binárias múltiplas é objeto de estudo de Zhang (1998), que considera como critérios de partição das amostras estatísticas baseadas nas matrizes de covariâncias amostrais, além de uma proposta paramétrica fundamentada em medida de entropia. O autor destaca a utilização da entropia como responsável pela produção de modelos mais estáveis e predições mais consistentes.

De'Ath (2002) propõe a construção de árvores de regressão multivariadas, aplicadas ao estudo da abundância de 12 espécies de aranhas caçadoras em dunas holandesas. O objetivo é avaliar quais fatores ambientais melhor explicam as variações dessas abundâncias na área sob estudo, bem como correlacionar as ocorrências das diferentes espécies. A exploração dos resultados produzidos pelo modelo multivariado é realizada por meio da construção de *biplots* (GOWER, 1996).

Miller e Franklin (2002) utilizam árvores de classificação com fins preditivos no estudo da distribuição espacial de quatro alianças de vegetação no deserto de Mojave, na Califórnia. Nesse estudo, são consideradas covariáveis ambientais como temperatura, precipitação e radiação solar, além de covariáveis relativas à localização das áreas sob estudo. Larsen e Speckman (2004), por sua vez, aplicam árvores de regressão multivariadas ao estudo da abundância de diversas espécies vegetais presentes na floresta de Ozark, Missouri.

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). Devido a sua flexibilidade, o autor batiza a técnica como *árvore de decisão generalizada multivariada*. Ressalta, no entanto, que a escolha adequada de uma estrutura de covariâncias para as respostas é uma etapa delicada da análise, devendo ser avaliada com bastante cuidado.

Como ressaltado anteriormente, a construção de árvores de regressão multivariadas requer critérios de partição e medidas de qualidade do ajuste adequadas ao estudo de variáveis respostas múltiplas. Para árvores de regressão multivariadas, por exemplo, De'Ath (2002) propõe a extensão do índice ANOVA considerando a soma dos desvios quadráticos dos vetores de observações em torno do vetor de médias dos nós. Propõe ainda medidas baseadas na soma dos desvios absolutos em torno da mediana e na soma das distâncias entre vetores de observações. Tais propostas são apresentadas de maneira mais detalhada a seguir.

Seja t um nó constituído por n elementos $\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_n}$, em que $\mathbf{y_j} = \left(y_{1j}, y_{2j}, ..., y_{qj}\right)$ é o vetor de respostas quantitativas associado ao elemento j, com vetores de médias e medianas $\bar{\mathbf{y}}_t$ e $\tilde{\mathbf{y}}_t$, respectivamente. Considere ainda $\mathbf{y_j^*} = \left(y_{1j}^*, y_{2j}^*, ..., y_{qj}^*\right)$ um vetor de respostas associado a uma observação independente daquelas utilizadas na construção do modelo e $d_{jj'}$ e d_j^* alguma medida de distância entre duas observações j e j' pertencentes a um mesmo nó e entre j e uma nova observação j^*, observação independente alocada ao nó em que j está presente. A Tabela 1 apresenta os critérios de partição e as medidas de erro de predição (para fins de validação cruzada) propostas por DeAth (2002).

Tabela 1 - Medidas de heterogeneidade e de erro de predição para árvores de regressão multivariadas propostas em DeAth (2002)

Descrição	Impureza	Erro de predição
Soma multivariada dos quadrados dos desvios em relação à média.	$\sum_{i,j}\left(y_{ij} - \bar{\mathbf{y}}_\mathbf{t}\right)^2$	$\sum_{j}\left(y^* - \bar{\mathbf{y}}_\mathbf{t}\right)^2$
Soma multivariada dos desvios absolutos em relação à mediana.	$\sum_{i,j}\left\lvert y_{ij} - \tilde{\mathbf{y}}_\mathbf{t}\right\rvert$	$\sum_{j}\left\lvert y^* - \tilde{\mathbf{y}}_\mathbf{t}\right\rvert$
Soma das distância quadráticas	$\sum_{j>j',j'} d_{jj'}^2$	$\sum_{j}\frac{d_j^{*2}}{n} - \sum_{j>j',j'}\frac{d_{jj'}^2}{n^2}$

2.3 Coeficientes de similaridade e dissimilaridade para múltiplas variáveis categorizadas

Uma das propostas apresentadas neste trabalho é a construção de árvores de classificação multivariadas com base em coeficientes de dissimilaridade. As seções seguintes apresentam diferentes coeficientes de similaridade e dissimilaridade para dados categorizados.

2.3.1 Coeficientes de similaridade e dissimilaridade para múltiplas variáveis binárias

Considere a comparação de um par de elementos (j e j') a partir dos resultados de q variáveis binárias, cada uma codificada de tal forma que possa assumir valores 0 ou 1 (por exemplo, 0 no caso de ausência de determinado sintoma clínico e 1 em sua presença). Dessa

forma, para cada variável, uma das seguintes configurações deve ser observada: 0-0, 0-1, 1-0 ou 1-1, sendo o primeiro valor relativo à observação j e o segundo à observação j'.

O conjunto de pares de resultados das *q* variáveis pode ser tabelado como apresentado a seguir:

Tabela 2 – Distribuição dos resultados de *q* atributos avaliados em um par de elementos

		Elemento j'		Total
		1	0	
Elemento j	1	a	b	$a+b$
	0	c	d	$c+d$
	Total	$a+c$	$b+d$	$q=a+b+c+d$

Neste caso, *a*, *b*, *c* e *d* representam os números de variáveis em que se verificou cada uma das quatro possíveis combinações de resultados. Usando tais quantidades, diversos coeficientes de similaridade podem ser estabelecidos. Particularmente, a família de coeficientes de similaridade expressa em (1) é caracterizada conforme as freqüências associadas à dupla-presença (*a*) e à dupla ausência (*d*) são consideradas.

$$S_{jj'} = \frac{\lambda a + \delta d}{\lambda a + b + c + \delta d}, \text{ sendo } \lambda, \delta \geq 0 \qquad (1)$$

Os parâmetros λ e δ controlam, respectivamente, o peso aferido à dupla presença e à dupla ausência. Tomando $\lambda = \delta = 1$, por exemplo, tem-se o coeficiente de similaridade simples, que consiste basicamente na proporção de resultados coincidentes. Caso se deseje desprezar duplas ausências, toma-se δ=0. Qualquer coeficiente de similaridade pertencente à família apresentada em (1) pode tomar valores no intervalo $[0,1]$. Desse modo, é possível definir um correspondente coeficiente de dissimilaridade assumindo resultados no mesmo intervalo de valores como sendo $D_{jj'} = 1 - S_{jj'}$.

Considerar, ou não, a freqüência de resultados do tipo 0-0 é uma questão que deve ser decidida à luz dos objetivos do estudo. Em pesquisas de mercado, por exemplo, detectar a não

aceitação de clientes ou consumidores com respeito a serviços ou produtos, pode ser tão ou mais relevante que a aceitação dos mesmos. Em estudos ecológicos, por outro lado, pode ser incoerente considerar similares duas locações distintas pelo fato de alguma espécie animal ou vegetal não ser encontrada em ambas. Cox e Cox (2001) apresentam diversos outros coeficientes de similaridade, destacando suas principais características.

2.3.2 Coeficiente de dissimilaridade simples para múltiplas variáveis categorizadas

Considere agora a situação em que se têm variáveis categorizadas com mais de duas categorias. A maneira mais simples de se estabelecer a dissimilaridade entre dois elementos é pela proporção de atributos não coincidentes. Suponha que dois elementos j e j' tenham sido avaliados em relação à q variáveis categorizadas. Para uma variável i, sejam k_i e k_i' os resultados desta variável avaliados em j e j', e $D_{k_j k_{j'}}$ uma variável indicadora definida como:

$$D_{k_i k_i'} = \begin{cases} 1, k_i \neq k_{i'} \\ 0, k_i = k_{i'} \end{cases}.$$

A dissimilaridade entre j e j' é quantificada pela proporção de resultados não coincidentes, ou seja,

$$D_{jj'} = \frac{1}{q}\sum_{i=1}^{q} D_{k_i k_{i'}}, \qquad (2)$$

podendo-se definir o correspondente coeficiente de similaridade como $S_{jj'} = 1 - D_{jj'}$.

O coeficiente de dissimilaridade apresentado em (2) destaca-se por sua simplicidade. Entretanto, não considera aspectos importantes relativos às variáveis sob estudo, como o número de possíveis resultados para cada variável, as taxas de ocorrências de observações em cada categoria e as associações existentes entre variáveis. Coeficientes de dissimilaridade mais elaborados permitem considerar um ou mais dos aspectos citados ao quantificar a similaridade entre dois elementos. Dois desses coeficientes são apresentados e discutidos nas seções seguintes.

2.3.3 Coeficiente de dissimilaridade baseado em probabilidades

Goodall (1966) propõe um coeficiente de similaridade baseado em probabilidades, o qual incorpora as incidências dos atributos categorizados sob estudo. Para calculá-lo, pares de

resultados de uma variável categorizada devem ser ordenados conforme suas respectivas similaridades. Pares de resultados não coincidentes são tomados como igualmente dissimilares. Pares de resultados coincidentes são considerados tão mais similares quanto menos prováveis as ocorrências de tais resultados. Uma breve descrição do coeficiente de similaridade proposto em Goodall (1966) é apresentada a seguir.

Seja Y_i uma variável categorizada com r_i possíveis resultados, sendo $p_k = P(Y_i = y_k)$, $k = 1,2,...,r_i$, as probabilidades de ocorrência de cada resultado. Goodall (1966) define inicialmente que a similaridade $S_{kk'}$ entre dois resultados k e k' de Y_i, é tal que $S_{kk'} = 0$, se $k \neq k'$, e $S_{kk'} > 0$, se $k = k'$. Além disso, o coeficiente assume valores maiores para coincidências de resultados menos prováveis. Suponha dois pares de observações de Y_i, $k = k'$ e $k^* = k^{*'}$, tais que $p_k < p_{k^*}$. Então $S_{kk'} > S_{k^*k^{*'}}$.

Seja $P_{kk'}$ a probabilidade de um par de elementos selecionados ao acaso apresentar similaridade igual ou superior a $S_{kk'}$. Tem-se:

$$P_{kk'} = \sum_{q \in Q} p_q^2, \; k = k',$$

sendo

$$Q = \left\{ k^* : \left(p_{k*} \leq p_k \right) \right\}, \tag{3}$$

e $P_{kk'} = 1$, se $k \neq k'$. Define-se $D_{kk'} = P_{kk'}$ como a medida de dissimilaridade entre k e k' e $S_{kk'} = 1 - P_{kk'}$ a correspondente medida de similaridade.

Na prática, as probabilidades p_k, $k = 1,2,...,r$ são desconhecidas, devendo ser aproximadas pelas freqüências (f_k) avaliadas em uma amostra de tamanho n disponível, fornecendo a seguinte estimativa para $D_{kk'}$:

$$\hat{D}_{kk'} = \begin{cases} \sum_{k^* \in Q} \dfrac{f_{k^*}\left(f_{k^*} - 1\right)}{n(n-1)}, & se\ k = k'; \\ 0 & , se\ k \neq k' \end{cases}$$

sendo Q determinado como em (3), mas com base nas freqüências amostrais.

Considere agora resultados relativos a dois elementos j e j', representados pelos vetores $\mathbf{y_j} = \left(y_{1j}, y_{2j}, ..., y_{qj}\right)$ e $\mathbf{y_{j'}} = \left(y_{1j'}, y_{2j'}, ..., y_{qj'}\right)$. Para uma variável Y_i ($i = 1,2,...,r_i$), suponha $\left\{Y_{i_1}, Y_{i_2}, ..., Y_{i_{ri}}\right\}$ o conjunto de seus resultados possíveis, com probabilidades de ocorrência p_{i_v}, $v = 1,2,...,r_i$. Define-se $P_{j_i j_i'}$ a probabilidade de que um par aleatório de resultados da variável i seja tão ou mais similar que o par de resultados j e j'. Considere, agora, dois pares de indivíduos (j e j'; h e h'). De maneira similar à estabelecida para pares de resultados de uma variável, o seguinte critério é estabelecido, agora com relação aos vetores de respostas:

$$\prod_{i=1}^{m} P_{j_i j_i'} < \prod_{i=1}^{m} P_{h_i h_i'} \rightarrow S_{jj'} > S_{hh'}.$$

A dissimilaridade $D_{jj'}$ é obtida pela soma das probabilidades associadas aos pares de vetores (h, h') tão ou mais similares que o par (j, j'), ou seja:

$$D_{jj'} = \sum_{h_1=1}^{r_1}\sum_{h_1'=1}^{r_1}\sum_{h_2=1}^{r_2}\sum_{h_2'=1}^{r_2}...\sum_{h_r=1}^{r_m}\sum_{h_r'=1}^{r_m}\prod_{i=1}^{r}\prod_{l=1}^{r} p_{h_i}\, p_{h_l'},$$

sendo que a soma é válida para o conjunto de vetores (h, h') tal que $\prod_{i=1}^{r} P_{h_i h_i'} \leq \prod_{i=1}^{r} P_{j_i j_i'}$. O correspondente coeficiente de similaridade é dado por $S_{jj'} = 1 - D_{jj'}$. Novamente, deve-se ressaltar que, frente ao desconhecimento das probabilidades inerentes ao cálculo do coeficiente, tais quantidades devem ser aproximadas pelas respectivas freqüências relativas, calculadas com base na amostra coletada.

Características como diferenças quanto ao número de categorias das variáveis e presença de resultados pouco freqüentes são tratadas de forma diferenciada pelo coeficiente descrito nesta seção. Sua utilização, no entanto, requer cuidado, uma vez que as associações entre variáveis não são incorporadas, inviabilizando a aplicação do coeficiente quando as variáveis consideradas são correlacionadas. Em situações deste tipo, recomenda-se, por exemplo, o coeficiente baseado em distribuições condicionais de probabilidades, proposto em Quang e Bao (2005) e apresentado na seção seguinte.

2.3.4 Coeficiente de dissimilaridade baseado em distribuições condicionais de probabilidades

Quang e Bao (2005) propõem um método "indireto" aplicado ao cálculo da dissimilaridade entre atributos categorizados, e, conseqüentemente, a pares de observações constituídas de múltiplos atributos desta natureza. O termo "indireto" refere-se ao fato de que o cálculo do coeficiente, quando aplicado aos resultados de uma determinada variável, baseia-se na distribuição condicional de probabilidades das demais variáveis em relação ao seu resultado. A utilização de distribuições condicionais no cálculo do coeficiente de dissimilaridade automaticamente incorpora as 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 de medidas de divergência entre as distribuições de probabilidades das demais variáveis condicionadas nos resultados considerados, ou seja,

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

sendo $cpd(\cdot)$ a distribuição de probabilidades condicionais e $\Psi(\cdot,\cdot)$ uma medida de divergência entre as duas distribuições de probabilidades. Quang e Bao (2005) consideram, com tal finalidade, o método de divergência de Kullback-Leibler (KULLBACK e LEIBER, 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 calculadas individualmente para cada variável:

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

Por incorporar dependências entre variáveis, seu uso é indicado caso as associações entre as variáveis consideradas não sejam nulas. Sugere-se quantificar tais associações a fim de justificar (ou não) a utilização desse coeficiente de dissimilaridade.

2.4 Medida de entropia para múltiplas variáveis

Como discutido na seção 2.1, medidas de entropia podem ser aplicadas para quantificar a heterogeneidade dos nós em árvores de classificação univariadas (BREIMAN et al., 1984). A utilização de uma extensão multivariada da medida de entropia como medida de heterogeneidade dos nós pode ser adequada, por considerar unicamente as distribuições de freqüências dos resultados das variáveis nos nós constituídos. O uso da entropia é freqüente, por exemplo, como medida de impureza dos grupos produzidos por análises de agrupamentos (LI E DARCY, 1980).

Propõe-se aqui considerar a medida de entropia como mais uma alternativa para a construção e seleção de árvores de classificação multivariadas. Sejam $Y_1, Y_2, ..., Y_q$ variáveis aleatórias qualitativas, cada uma com r_i possíveis resultados ($i = 1,2,...,q$). A entropia do vetor aleatório $\mathbf{Y} = (Y_1, Y_2, ..., Y_q)$ é definida como:

$$H(\mathbf{Y}) = \sum_{i=1}^{q} H(Y_i),$$

sendo $H(Y_i) = -\sum_{k=1}^{r_i} [P(Y_i = y_k)] \log_2 [P(Y_i = y_k)]$.

Dada uma amostra ς, a entropia multivariada pode então ser estimada por meio de

$$H(\mathbf{Y} / \varsigma) = \sum_{i=1}^{q} H(Y_i \mid \varsigma), \qquad (4)$$

em que $H(Y_i \mid \varsigma) = -\sum_{k=1}^{r_i} [p_{ik}] \log_2 [p_{ik}]$, sendo p_{ik} a proporção amostral de resultados k da variável i. Novamente, caso $p_{ik} = 0$, $\forall k \in \{1,2,...,r\}$, considera-se $[p_{ik}] \log_2 [p_{ik}] = 0$.

O valor de $H(Y_i)$ está atrelado a r_i, o número de atributos da *i*-ésima variável, podendo assumir valores no intervalo $[0, \log_2 r_i]$. Assim, ao considerar conjuntamente variáveis com diferentes números de categorias, o valor da medida de entropia multivariada é influenciado de maneira mais acentuada por variáveis com maior gama de possíveis resultados. A fim de evitar

distorções ocasionadas pelas diferentes amplitudes dos coeficientes de entropia, pode-se padronizar $H(Y_i)$ da seguinte maneira:

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

garantindo valores de $H^*(Y_i)$ no intervalo $[0,1]$, para qualquer r_i. Pode-se estabelecer uma medida de entropia conjunta para $Y_1, Y_2, ..., Y_q$ de maneira similar à apresentada em (4), obtendo-se:

$$H^*(\mathbf{Y} \mid \varsigma) = \frac{1}{q}\sum_{i=1}^{q} H^*(Y_i \mid \varsigma), \tag{5}$$

sendo que a divisão por q garante valores de $H^*(\mathbf{Y} \mid \varsigma)$ no intervalo $[0,1]$.

2.5 Geração de valores amostrais para a distribuição multinomial e para distribuições multivariadas de probabilidades

Busca-se, neste trabalho, desenvolver metodologias de classificação por árvores capazes de explicar a variação de um conjunto de variáveis categorizadas com base em resultados de covariáveis. Nesse contexto, torna-se importante avaliar o desempenho dessas metodologias quando aplicadas na explicação de variáveis aleatórias multinomiais com diferentes graus de associação. Com este objetivo, um estudo por simulação é proposto, tornando necessária a geração de variáveis multinomiais dependentes. As seções seguintes apresentam algoritmos adequados à geração de resultados de uma variável aleatória multinomial e de distribuições de probabilidades multivariadas. A conjunção desses dois algoritmos possibilita, conforme o desejado, gerar valores amostrais para múltiplas variáveis multinomiais.

2.5.1 Geração de valores amostrais para uma variável aleatória com distribuição multinomial

A simulação de variáveis aleatórias com distribuição multinomial pode ser realizada por meio de um algoritmo de busca seqüencial (BUSTOS E ORGAMBIDE, 1992). Suponha Y uma variável multinomial com r categorias ($y_1, y_2, \ldots, y_r$), tal que $P(Y = y_k) = p_k$ e $\sum_{k=1}^{r} p_k = 1$. A

geração de n resultados da variável Y, segundo o método de busca seqüencial, é descrita pelo algoritmo 1.

Algoritmo 1

i. Gerar $u \sim U(0,1)$ segundo algum dos diversos algoritmos disponíveis para simulação de variáveis aleatórias com distribuição uniforme (BUSTOS, ORGAMBIDE, 1992);
ii. Atribuir $k = 0$; $s = 0$;
iii. Enquanto $u > s$, fazer $k = k + 1$, $s = s + p_k$;
iv. Retornar y_k;
v. Repetir os passos $(i - iv)$ n vezes.

2.5.2 Geração de valores amostrais para variáveis aleatórias multivariadas correlacionadas

A geração de valores amostrais para um par de variáveis aleatórias correlacionadas pode ser realizada a partir de resultados de um par de variáveis correlacionadas uniformemente distribuídas no intervalo [0,1] (Dias, 1996). A geração de distribuições uniformes bivariadas, por sua vez, pode ser executada com base nas funções de distribuição de probabilidades acumuladas marginais de um par de variáveis aleatórias contínuas (Z_1 , Z_2) que apresentem o coeficiente de correlação desejado. Uma alternativa consiste em simular variáveis aleatórias normais correlacionadas, dada a disponibilidade de tal procedimento em um grande número de programas estatísticos, como, por exemplo, no software R (R DEVELOPMENT CORE TEAM, 2007). O algoritmo 2 descreve o procedimento aplicado à geração de n valores amostrais para um vetor aleatório $\mathbf{Z} = (Z_1 , Z_2)$, normalmente distribuído, com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$.

Algoritmo 2

i. Gerar n valores amostrais para o vetor aleatório $\mathbf{Z}^* = (Z_1^*, Z_2^*)$, normalmente distribuído, com média $\boldsymbol{\mu}_{Z^*} = \mathbf{0}$ e matriz de covariâncias $\boldsymbol{\Sigma}_{\mathbf{Z}^*} = \mathbf{I}_2$, resultando em uma matriz $\mathbf{Z}^*$, com

dimensão $n \times 2$. A geração de Z_1^* e Z_2^* pode ser efetuada pelo método da distribuição inversa (BUSTOS E ORGAMBIDE, 1992);

ii. Mediante decomposição espectral de $\mathbf{\Sigma}$, obter as matrizes

$$\mathbf{A} = \begin{bmatrix} \mathbf{a_1} \\ \mathbf{a_2} \end{bmatrix} ; \mathbf{D} = \begin{bmatrix} \sqrt{d_1} & 0 \\ 0 & \sqrt{d_2} \end{bmatrix},$$

sendo $(\mathbf{a_i}, d_i), i = 1,2$, autovetores e autovalores de $\mathbf{\Sigma}$;

iii. Calcular $\mathbf{Z} = \boldsymbol{\mu} + \mathbf{ADZ}^{*\prime}$.

Seja $F_{\mathbf{Z}}$ a distribuição de probabilidades conjunta de $\mathbf{Z}$, F_{Z_1} e F_{Z_2} as funções de distribuição marginais e ρ_{Z_1,Z_2} a correlação entre Z_1 e Z_2. Então, segundo Dias (1996), as variáveis $U_1 = F_{Z_1}$ e $U_2 = F_{Z_2}$ têm distribuição uniforme no intervalo (0,1), com coeficiente de correlação ρ_{U_1,U_2}. Se Z_1 e Z_2 são independentes, U_1 e U_2 também o são, e caso contrário, as variáveis aleatórias uniformes refletem a correlação das variáveis aleatórias originais. A geração de um par de variáveis aleatórias correlacionadas Y_1 e Y_2 com distribuições marginais de probabilidades F_{Y_1} e F_{Y_2} pode ser realizada, por exemplo, utilizando o método da distribuição inversa (BUSTOS E ORGAMBIDE, 1992), desde que ambas F_{Y_1} e F_{Y_2} sejam invertíveis, conforme descrito no algoritmo 3.

Algoritmo 3

i. Gerar $\mathbf{z} = (z_1, z_2)$ a partir de $F_{\mathbf{Z}}$, uma distribuição de probabilidades normal bivariada, com parâmetros

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} ; \mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix};$$

ii. Calcular $u_1 = F_{Z_1}(z_1), u_2 = F_{Z_2}(z_2)$;

iii. Calcular $y_1 = F_{Y_1}^{-1}(u_1), y_2 = F_{Y_2}^{-1}(u_2)$.

iv. Repetir os passos anteriores n vezes.

Dias (1996) apresenta resultados obtidos por simulação que atestam a correspondência entre os coeficientes de correlação das variáveis aleatórias originais, normalmente distribuídas, e as correlações das variáveis produzidas conforme apresentado no algoritmo 3, abordando distribuições de probabilidades como a exponencial, a triangular e a binomial. Dado o interesse em gerar variáveis multinomiais dependentes, avalia-se neste trabalho a possibilidade de extensão do método de busca seqüencial para o caso multivariado, mediante execução de um algoritmo resultante da composição dos algoritmos 1 e 3.

2.6 Análise de correspondência

A análise de correspondência (RENCHER, 1995) é uma técnica multivariada que possibilita a representação das categorias que compõem tabelas de contingência com duas ou mais entradas em gráficos bidimensionais, de tal forma que essa representação consiga refletir uma expressiva parcela da variação original dos dados. O gráfico produzido pela análise de correspondência é um diagrama de dispersão, em que as categorias das variáveis sob estudo são representadas por pontos. A associação entre categorias pode ser avaliada com base nas proximidades entre pontos: pontos próximos indicam que as respectivas categorias, em conjunto, apresentam freqüência maior do que aquela esperada sob a hipótese de independência das variáveis, apontando associação direta entre as mesmas. Pontos distantes, por sua vez, indicam que a freqüência associada às respectivas categorias está abaixo do esperado sob independência, apontando associação inversa entre as mesmas. A técnica é denominada análise de correspondência múltipla quando três ou mais variáveis são analisadas simultaneamente.

3 METODOLOGIA

O presente capítulo divide-se em duas seções: a Seção 3.1 descreve o conjunto de dados analisado neste trabalho, enquanto a Seção 3.2 introduz as metodologias originais de árvores de classificação multivariadas.

3.1 Material

As metodologias originais de classificação multivariadas por árvores são 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 a Comunidade Européia, o Instituto Norte Americano de Alcoolismo e Abuso do Álcool, a Agência de Educação e Ciência da Suiça, o Ministério de Saúde Pública da Alemanha, além de 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 elevado de bebidas alcoólicas 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. A seleção da amostra foi realizada via amostragem estratificada, levando em consideração a representatividade de diferentes níveis educacionais e sócio-econômicos na composição do município de Botucatu. Somente indivíduos com mais de 17 anos estavam aptos a serem entrevistados. Cada estrato era formado por setores censitários e os entrevistados foram selecionados conforme delineamento amostral por conglomerados. Consideraram-se como unidades amostrais casas de famílias, excluindo-se, por exemplo, moradias estudantis e estabelecimentos comerciais. Mais do que uma pessoa poderia ser entrevistada por domicílio. Aproximadamente 5,8% dos indivíduos selecionados recusaram-se a participar do estudo. A Tabela 3 apresenta as variáveis consideradas na aplicação dos modelos propostos, suas codificações e classificações.

Tabela 3 – Variáveis relativas ao consumo alcoólico, cigarro e maconha, além de características pessoais, componentes 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
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
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.

Tabela 3 – Variáveis relativas ao consumo alcoólico, cigarro e maconha, além de características pessoais, componentes de questionário aplicado a uma amostra de habitantes do município de Botucatu (SP) (conclusão)

Variável	Descrição	Tipo
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
CONSUMO*	Freqüência do consumo alcoólico nos últimos 12 meses.	1 – Nenhuma 2 – Poucas vezes (menos de 12 vezes) 3 – Muitas vezes (ao menos uma vez por mês)
INTENSIDADE*	Intensidade com que consumiu álcool num único dia, quando mais bebeu nos últimos 12 meses**.	1 – Nada; 2 – Menos de cinco drinques; 3 – Cinco drinques ou mais.
PREFERÊNCIA*	Bebida alcoólica preferida.	1 – Nenhuma; 2 – Cerveja; 3 – Vinho; 4 – Destilado.
CIGARRO*	Faz uso de cigarro.	1 – Sim; 2 – Não.
MACONHA*	Faz uso de maconha.	1 – Sim; 2 – Não.

* Variável resposta.

* *A medida de consumo alcoólico é ponderada conforme os números de doses e os tipos de bebidas citados pelo entrevistado.

3.2 Métodos

Nesta seção, são apresentadas as metodologias originais aplicadas à construção de árvores de classificação multivariadas.

3.2.1 Geração de valores amostrais para variáveis multinomiais dependentes

Como dito anteriormente, considera-se no presente trabalho utilizar conjuntamente os algoritmos 1 e 3, apresentados nas Seções 2.5.1 e 2.5.2, com o objetivo de gerar valores amostrais para múltiplas variáveis multinomiais. O algoritmo 4, resultado da junção dos dois algoritmos mencionados, é apresentado na seqüência.

Sejam Y_1 e Y_2 variáveis multinomiais com r_1 e r_2 categorias, respectivamente, tais que $P(Y_i = y_{k_i}) = p_{k_i}$, $\sum_{k_i=1}^{r_i} p_{k_i} = 1$, $i = 1,2$. A geração de n valores amostrais para $\mathbf{Y} = (Y_1, Y_2)$ é realizada da seguinte forma.

Algoritmo 4

i. Gerar um resultado amostral para o vetor aleatório $\mathbf{Z} = (Z_1, Z_2)$, normalmente distribuído com vetor de médias e matriz de covariâncias dados, respectivamente, por:

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \; ; \; \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \sigma_{Z_1,Z_2} \\ \sigma_{Z_1,Z_2} & 1 \end{bmatrix},$$

sendo $\sigma_{Z_1,Z_2} = \rho_{Z_1,Z_2}$, o coeficiente de correlação linear de Pearson;

ii. Calcular $u_1 = F_{Z_1}(z_1), u_2 = F_{Z_2}(z_2)$;

iii. Atribuir $k_i = 0$, $s_i = 0$, $i = 1,2$;

iv. Enquanto $u_i > s_i$, fazer $k_i = k_i + 1$, $s_{k_i} = s_{k_i} + p_{k_i}$, $i = 1,2$;

v. Repetir os passos anteriores n vezes.

A geração de $q > 2$ variáveis aleatórias multinomiais a partir de q variáveis normais correlacionadas se dá com a extensão do procedimento apresentado no algoritmo 4. Esse algoritmo tem seu desempenho avaliado na geração de pares de variáveis com diferentes níveis de associação, números de categorias e graus de entropia, com base em um estudo por simulação.

Foram consideradas variáveis aleatórias multinomiais com $r = 3, 4, 6$ e 8 categorias. Os vetores de probabilidades, para cada valor de r, foram selecionados mediante o processo descrito no algoritmo 5, visando considerar variáveis com cinco graus distintos de entropia.

Algoritmo 5

i. Simulação de um vetor de probabilidades $\mathbf{p} = (p_1, p_2, ..., p_r), \sum_{k=1}^{r} p_k = 1$. Obtém-se p_k gerando um resultado para U, variável aleatória com distribuição uniforme no intervalo $\left[0, 1 - \sum_{k'=1}^{k-1} p_{k'}\right]$, se $k > 1$, ou no intervalo $[0, 1]$, se $k = 1$;

ii. Cálculo do valor da medida de entropia para $\mathbf{p}$:

$$\phi(\mathbf{p}) = -\sum_{k=1}^{r} p_k \log_2 p_k ;$$

iii. Repetição dos passos *a* e *b* 1000 vezes, produzindo uma seqüência de vetores $\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_{1000}$, com respectivas entropias $\phi(\mathbf{p}_1), \phi(\mathbf{p}_2), ..., \phi(\mathbf{p}_{1000})$.

iv. Seleção dos vetores de probabilidades $\mathbf{p}_{(10\%)}$, $\mathbf{p}_{(25\%)}$, $\mathbf{p}_{(50\%)}$, $\mathbf{p}_{(75\%)}$, $\mathbf{p}_{(90\%)}$ associados aos quantis $10\%, 25\%, 50\%, 75\%$ e 90% do conjunto de 1000 entropias obtidas.

A partir dos vetores de probabilidades selecionados, foram simulados 1000 valores amostrais de $\mathbf{Y} = (Y_1, Y_2)$, variáveis aleatórias multinomiais dependentes, originadas por variáveis aleatórias normais correlacionadas, com coeficiente de correlação ρ, gerado aleatoriamente no intervalo $(-1, 1)$. Calculou-se, como medida de associação entre as variáveis Y_1 e Y_2 o valor da estatística χ^2 (AGRESTI, 2000), que mede o afastamento da distribuição de freqüências conjunta observada em relação às freqüências esperadas sob a hipótese de não associação.

Para se avaliar a correspondência entre a correlação das variáveis aleatórias normais originais e a associação das variáveis multinomiais, gráficos de dispersão de χ^2 em função de ρ foram obtidos por simulação. O algoritmo 6 delineia o estudo por simulação, aplicado a cada combinação de k, $\mathbf{p}_{(q_1)}$ e $\mathbf{p}_{(q_2)}$, $q_1, q_2 \in \{10\%, 25\%, 50\%, 75\%, 90\%\}$.

Algoritmo 6

i. Geração de ρ aleatoriamente no intervalo $(-1,1)$;

ii. Geração de 1000 resultados para o vetor aleatório $\mathbf{Z} = (Z_1, Z_2)$, com $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{\Sigma})$, sendo

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix};$$

iii. Geração de 1000 resultados para o vetor aleatório $\mathbf{Y} = (Y_1, Y_2)$, $Y_1 \sim Mult(1;\, p_{(q_i)})$, $Y_2 \sim Mult(1;\, p_{(q_2)})$, conforme descrito no algoritmo 3;

iv. Cálculo do valor da estatística χ^2 para as variáveis Y_1 e Y_2;

v. Repetição dos passos anteriores 1000 vezes, produzindo pares de resultados distintos para $\left\{\left(\rho_i, \chi_i^2\right),\, i = 1,2,\cdots,1000\right\}$;

vi. Construção de um gráfico formado pelos pontos de coordenadas $\left(\rho_1, \chi_1^2\right), \left(\rho_2, \chi_2^2\right), \cdots, \left(\rho_{1000}, \chi_{1000}^2\right)$.

3.2.2 Construção de árvores de classificação multivariadas com base em coeficientes de dissimilaridade

As etapas relativas à construção de árvores de classificação, apresentadas no Capítulo 2 sob o enfoque univariado, são agora adaptadas, permitindo considerar múltiplas variáveis respostas. Primeiramente, para uma medida de dissimilaridade D qualquer, seja $D_{jj'}$ a dissimilaridade calculada para dois elementos j e j'. Propõe-se como medida de heterogeneidade para n elementos que constituem um nó t a dissimilaridade média entre tais elementos, ou seja

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

Suponha que um nó t seja partido em dois novos nós t_L e t_R, com n_L e n_R elementos, respectivamente. A qualidade da partição executada é quantificada pela conseqüente variação da dissimilaridade média dentro dos nós, calculada conforme descrito a seguir.

$$\Delta_{\phi_{Dis}}(s,t)=\left[\frac{n(n-1)}{2}\right]^{-1}\sum_{j=1}^{n}\sum_{j'<j}D_{jj'}-\frac{n_L}{n}\left[\frac{n_L(n_L-1)}{2}\right]^{-1}\sum_{j=1}^{n_L}\sum_{j'<j}D_{jj'}-\frac{n_R}{n}\left[\frac{n_R(n_R-1)}{2}\right]^{-1}\sum_{j=1}^{n_R}\sum_{j'<j}D_{jj'} \quad (6)$$

Para uma árvore T, seja $\tilde{T}$ seu conjunto de nós finais. Toma-se como medida de heterogeneidade da árvore a dissimilaridade média entre pares de observações alocadas a um mesmo nó, ponderadas pelos tamanhos dos nós constituídos, conforme descrito na seqüência:

$$R_{Dis}(T)=\sum_{t\subset\tilde{T}}\frac{n_t}{n}\phi_{Dis}(t),$$

sendo n_t o número de elementos em t.

A construção do modelo multivariado de árvore de classificação se dá mediante execução de sucessivas partições binárias da amostra original, com base na medida de variação de similaridade apresentada em (6), até a obtenção de uma árvore $T_{\max}$, com poucos elementos em cada nó final. Obtida a árvore, deve-se iniciar a poda, conforme descrito em Breiman et al. (1984), desfazendo sucessivamente as partições, optando sempre por eliminar aquelas responsáveis por menores reduções da dissimilaridade ($\Delta_{\phi_{Dis}}(s,t)$), até que reste apenas um nó. O procedimento de poda permite a obtenção de uma seqüência aninhada de árvores, dentre as quais uma será selecionada com base em resultados alcançados via validação cruzada.

Breiman et al. (1984) ressalta a importância de se validar qualquer modelo constituído utilizando observações independentes daquelas empregadas em sua construção. Justifica argumentando que medidas de má-classificação calculadas a partir dos dados usados na construção das árvores tendem a produzir resultados muito otimistas com respeito à qualidade do ajuste. Sugere a utilização de amostras testes (caso se disponha de um grande número de elementos) ou validação cruzada. O procedimento de validação cruzada aplicado à construção de árvores de classificação multivariadas, baseado em coeficientes de similaridade, é apresentado na seqüência.

Seja T uma árvore de classificação multivariada. Suponha que uma nova observação y^*, independente daquelas utilizadas na construção de T, seja alocada ao nó t através da estrutura de classificação 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 entre esta nova observação e aquelas contidas em t, ou seja,

$$\phi_{Dis}\left(y^*\right) = \sum_{j \subset t} d_j^* / n_t \, .$$

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 $v = 1,2,...,V$. A taxa global de heterogeneidade da árvore, estimada por validação cruzada, é descrita como:

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

sendo $R\left(T^{(v)}\right) = \sum_{y_j \subset \varsigma_v} \frac{\phi_{Dis}\left(y_j\right)}{n_v}$ e n_v o número de elementos em ς_v. A seleção da melhor árvore é realizada então por meio da construção do gráfico de complexidade e aplicação da regra do desvio padrão, como descrito na Seção 2.1.

Propõe-se ainda neste trabalho um procedimento alternativo para seleção de modelos de árvores de classificação, também baseado na curva de custo-complexidade, porém sem levar em conta o desvio padrão da medida de qualidade do ajuste considerada. A Seção 3.2.4 apresenta de forma detalhada este novo critério. Sua aplicação nos exemplos simulados e na análise dos dados de consumo alcoólico (Seção 3.1) serve como referência para avaliar seu desempenho na busca de um modelo parcimonioso.

Outra etapa fundamental da análise de classificação e regressão por árvores é a definição de um critério de classificação dos nós finais. Seja $p_t(\mathbf{y})$ a distribuição de freqüências em um nó final t. A caracterização dos nós finais da árvore $\tilde{T}$ pode ser realizada segundo as freqüências conjuntas, classificando t por $\mathbf{y}=(y_1, y_2, ..., y_q)$ tal que $p_t(\mathbf{y})$ é máximo, ou segundo as freqüências individuais de cada variável, classificando um nó t por $\mathbf{y}=(y_1, y_2, ..., y_q)$ tal que $p_t(y_i)$ é máximo, $i = 1,2,...,q$.

3.2.3 Construção de árvores de classificação multivariadas com base em medida de entropia

Descreve-se nesta seção a proposta original de construção de árvores de classificação multivariadas baseadas no coeficiente de entropia. Considere $H^*(t) = H^*(\mathbf{Y} \mid t)$ o valor da entropia multivariada calculado a partir dos elementos presentes em um nó t, calculado com base nos resultados de um vetor de variáveis aleatórias categorizadas $\mathbf{Y}$, conforme descrito em (5). Suponha que uma partição s seja executada, produzindo dois novos nós, t_L e t_R. Propõe-se como critério para seleção da melhor partição aquela responsável por maximizar:

$$\Delta_{Ent}(s,t) = H^*(t) - \frac{n_L}{n} H^*(t) - \frac{n_R}{n} H^*(t).$$

A obtenção da seqüência de árvores aninhadas e conseqüente poda são realizadas de maneira semelhante à apresentada na Seção 3.2.2, mas com base nos resultados de $\Delta_{Ent}(s,t)$. Para uma árvore T com $\tilde{T}$ representando seu conjunto de nós finais, pode-se calcular a taxa global de impureza como

$$R_{Ent}(T) = \sum_{t \subset \tilde{T}} \frac{n_t}{n} H^*(y).$$

A seleção do modelo deve ser executada via validação cruzada, com base nos resultados de algum coeficiente de dissimilaridade, conforme descrito na Seção 3.2.2.

3.2.4 Seleção do modelo de árvore de classificação pelo ponto da curva de custo-complexidade mais afastado de uma reta representando variação uniforme do custo

Como critério alternativo à “regra do desvio padrão” (BREIMAN et al., 1984), propõe-se neste trabalho a escolha do melhor modelo pelo ponto do gráfico de custo-complexidade que mais se afasta de uma reta unindo os pontos referentes às arvores com maior e menor custo de má-classificação. Busca-se com isso determinar qual das árvores da seqüência aninhada minimiza, conjuntamente, custo e complexidade. A aplicação desse critério requer alguns cuidados, como descrito na seqüência.

Suponha $T_1, T_2, ..., T_{MAX}$ a seqüência de árvores aninhadas, $|\tilde{T}_1|, |\tilde{T}_2|, ..., |\tilde{T}_{MAX}|$ seus tamanhos e $R(T_1), R(T_2), ..., R(T_{MAX})$ as respectivas medidas de heterogeneidade. Calcula-se para cada árvore g:

$$|\tilde{T}_g|^* = \frac{|\tilde{T}_g| - m_{|\tilde{T}|}}{s_{|\tilde{T}|}} \;\; ; \;\; R^*(T_g) = \frac{R(T_g) - m_{R(T)}}{s_{R(T)}}$$

sendo m e s, os operadores média e desvio padrão. Seja d_g a distâncias do ponto da curva de custo-complexidade associado a uma árvore T_g da reta ligando os pontos inicial e final. Seleciona-se T_G tal que $d_G = \max_g(d_g)$. Como regras adicionais, caso $d_G = d_{G'} = \max(d_g)$ opta-se pela árvore de menor tamanho, ou seja, por T_G se $|\tilde{T}_G| < |\tilde{T}_{G'}|$, e por $T_{G'}$, caso contrário. Além disso, define-se $d_G = 0$ caso o ponto associado à árvore T_G esteja acima do ponto associado a T_1, situação em que não seria vantajoso considerar tal árvore como instrumento de classificação. Um dos méritos desse novo método de seleção de árvores de classificação multivariadas é que, ao contrário do critério do desvio padrão, ele não incorpora a estimativa do erro padrão da medida de heterogeneidade, resultando em menor esforço computacional. A Figura 2 ilustra a aplicação do método proposto para seleção de modelos.

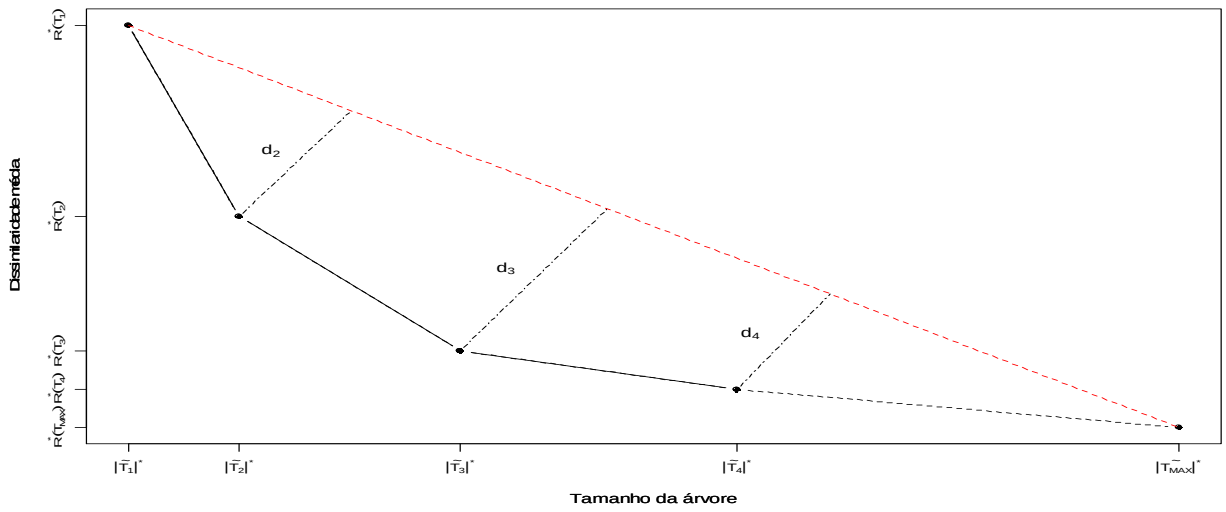

Figura 2 – Gráfico de custo complexidade com medidas padronizadas e distâncias dos pontos da curva de custo-complexidade à reta ligando os pontos inicial e final

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

Foram gerados $n = 500$ vetores de observações, compostos por três variáveis respostas multinomiais, cada uma delas com quatro categorias, e cinco covariáveis, com distribuições contínuas (normal e qui-quadrado) e discretas (Poisson e multinomiais) de probabilidades. As oito variáveis foram geradas a partir de um conjunto de oito variáveis normalmente distribuídas, com vetor de médias **0** e matriz de covariâncias $\boldsymbol{\Sigma}$, sendo os elementos da diagonal de $\boldsymbol{\Sigma}$ iguais a 1. Dessa forma, para qualquer par de variáveis, o valor da covariância equivale ao coeficiente de correlação linear de Pearson (Zar,1999). No presente estudo, quatro matrizes $\boldsymbol{\Sigma}$ foram consideradas, conforme a magnitude de seus componentes $\sigma_{ii'}, i \neq i'$:

- $\sigma_{ii'} = 0 , \forall\, i \neq i'$: covariâncias (e correlações) nulas;
- $|\sigma_{ii'}| \leq 0{,}5 , \forall\, i \neq i'$: covariâncias (e correlações) baixas;
- $|\sigma_{ii'}| \in [0{,}1] , \forall\, i \neq i'$: covariâncias (e correlações) variadas.
- $|\sigma_{ii'}| \geq 0{,}5 , \forall\, i \neq i'$: covariâncias (e correlações) elevadas;

Quanto às entropias das variáveis dependentes (Y_1, Y_2, Y_3), procedeu-se com a aplicação do algoritmo 5 e seleção dos vetores de parâmetros associados aos quantis 10% (entropia baixa), 50% (entropia moderada) e 90% (entropia alta) das entropias geradas. As seguintes configurações 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;
- Y_1, Y_2 e Y_3 geradas com entropias altas.

Sob cada uma das 12 configurações, resultantes de combinações de variáveis com diferentes graus de correlações e entropias, os dados gerados foram analisados mediante construção de árvores de classificação multivariadas baseadas em coeficientes de dissimilaridade e na medida de entropia. Os procedimentos propostos para construção de árvores de classificação

multivariadas têm seus desempenhos avaliados com base em requisitos como suas capacidades de discriminação, relativa à homogeneidade dos elementos que compõem cada nó final originado, e preditiva, referente à habilidade do modelo em predizer corretamente elementos independentes daqueles utilizados em sua construção. A capacidade de discriminação é avaliada por meio do cálculo da entropia ou da dissimilaridade média (estimada por validação cruzada) dos nós finais do modelo. A capacidade preditiva é avaliada com base na proporção de predições corretas, estimada também por validação cruzada.

A exploração dos resultados produzidos é realizada mediante construção de gráficos e aplicação de análise de correspondência múltipla, com o objetivo de investigar e confirmar a associação entre os nós constituídos e o conjunto de variáveis respostas. A análise de correspondência é realizada considerando o conjunto de variáveis respostas e uma nova variável qualitativa, indicando o nó ao qual cada indivíduo é alocado segundo a árvore construída. O software estatístico R (R DEVELOPMENT CORE TEAM, 2007) foi utilizado em todas as etapas deste trabalho, desde a programação dos algoritmos, execução do estudo por simulação e análise dos dados sobre consumo alcoólico dos habitantes do município de Botucatu.

4 RESULTADOS

4.1 Geração de valores amostrais para múltiplas variáveis aleatórias multinomiais

Antes de apresentar os resultados relativos à geração de valores amostrais para múltiplas variáveis multinomiais, a Figura 3 ilustra gráficos de densidades não paramétricas para as entropias geradas mediante execução do algoritmo 5. Percebe-se acentuada assimetria da distribuição das entropias geradas para variáveis com $r = 3$ e $r = 4$ categorias. Conforme se aumenta o número de categorias, a distribuição torna-se menos assimétrica.

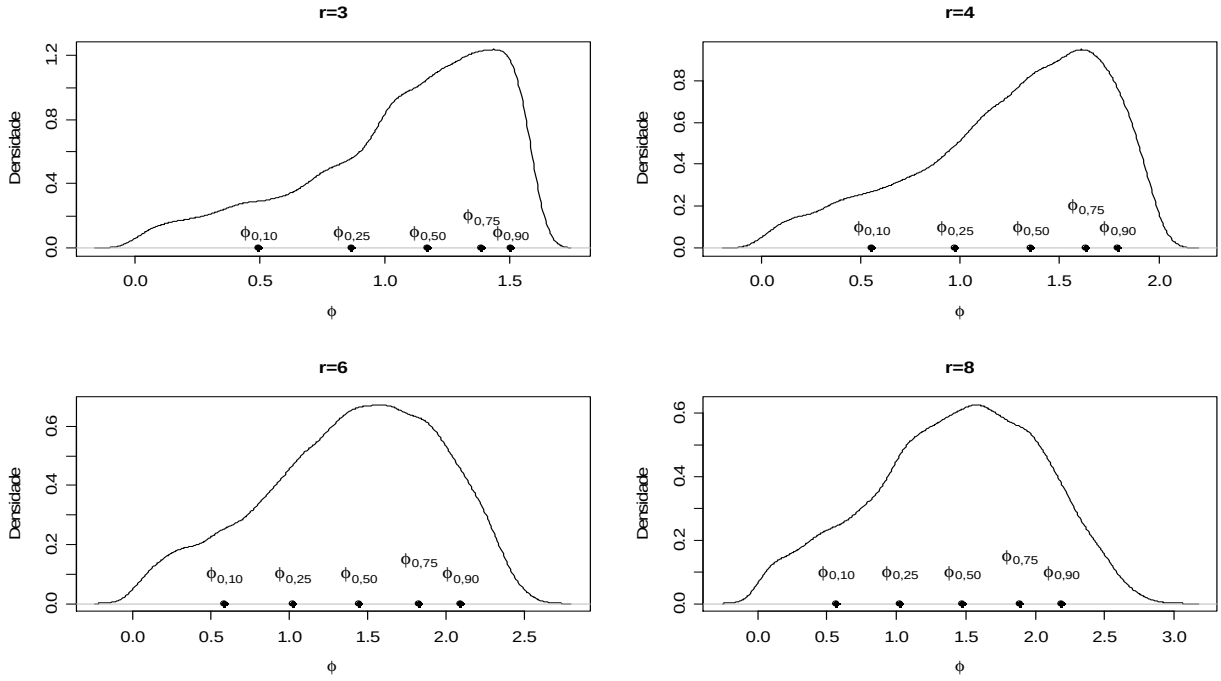

Figura 3 – Densidades não paramétricas empíricas de ϕ e respectivos quantis 10%, 25%, 50%, 75% e 90%, para variáveis com $r = 3, 4, 6$ e 8 categorias

As Figuras 4 a 7 apresentam os resultados do estudo por simulação proposto na Seção 3.2.1. Há uma maior correspondência entre a correlação das variáveis aleatórias normais originais e a associação entre as variáveis multinomiais produzidas à medida que aumentam as entropias das variáveis consideradas.

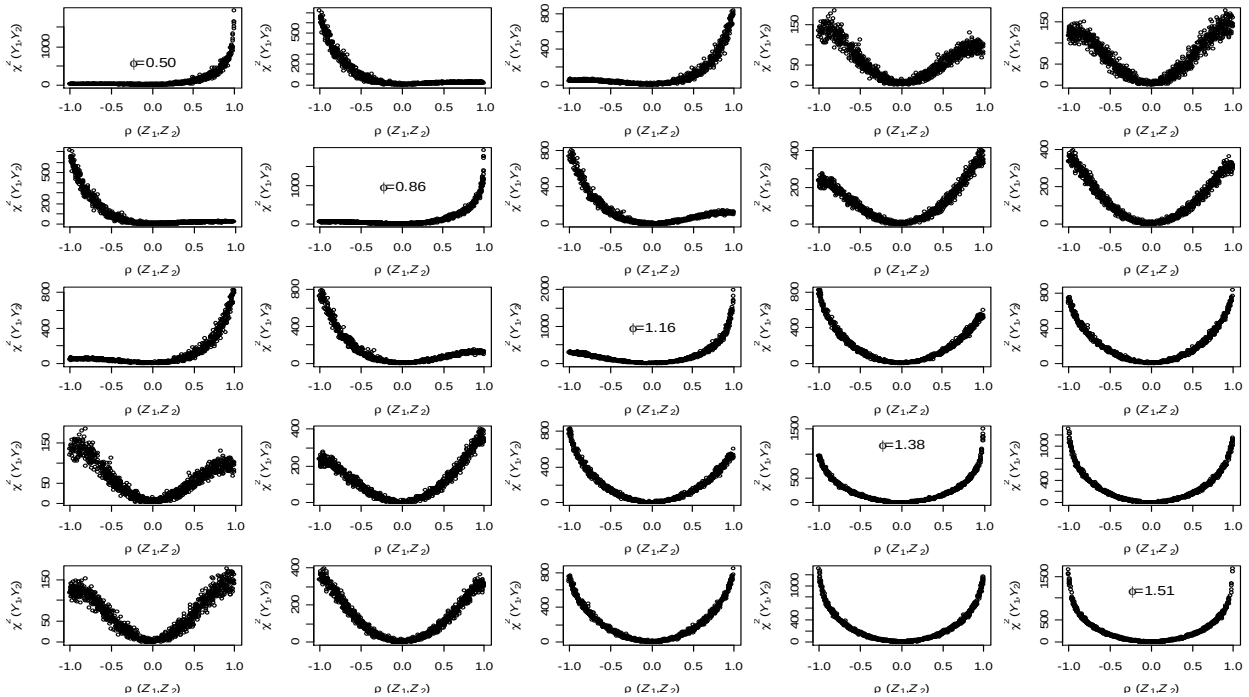

Figura 4 – Gráficos de dispersão para os coeficientes de correlação de Z_1 e Z_2, variáveis originais normalmente distribuídas, e os valores da estatística χ^2 para Y_1 e Y_2, variáveis multinomiais com $r = 3$ categorias, sob cinco diferentes graus de entropia (ϕ) indicados nos gráficos da diagonal

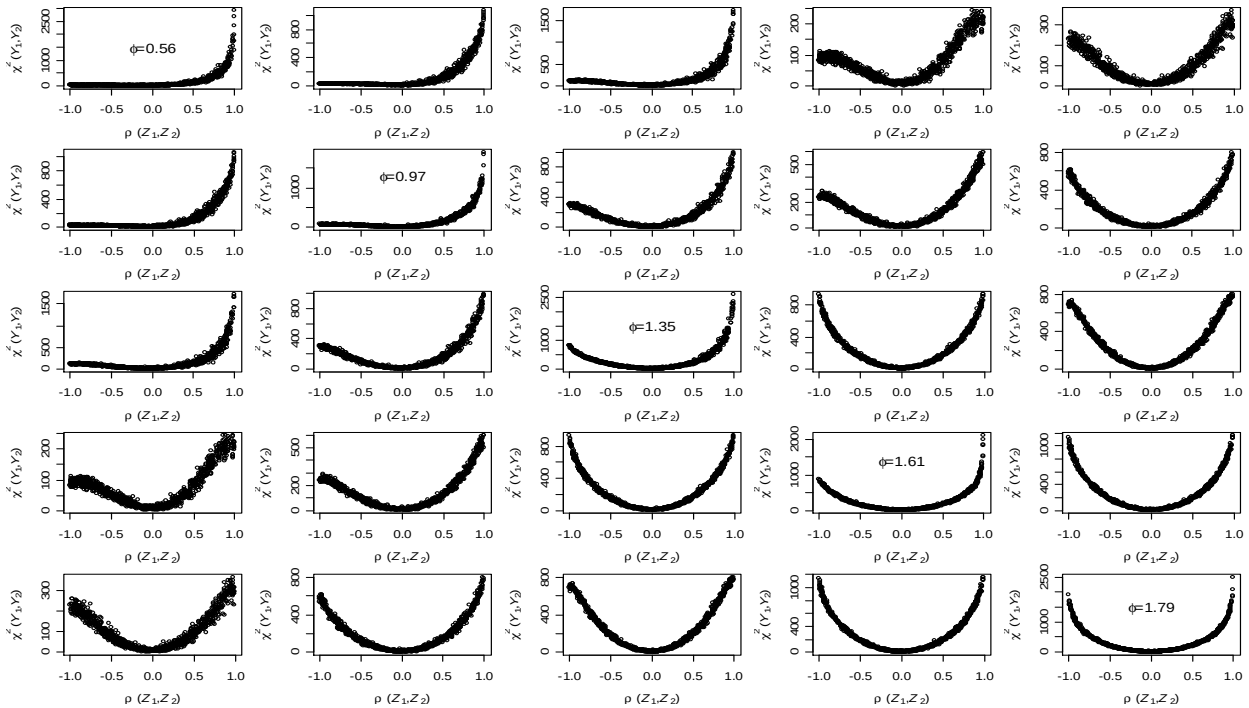

Figura 5 – Gráficos de dispersão para os coeficientes de correlação de Z_1 e Z_2, variáveis originais normalmente distribuídas, e os valores da estatística χ^2 para Y_1 e Y_2, variáveis multinomiais com $r = 4$ categorias, sob cinco diferentes graus de entropia (ϕ) indicados nos gráficos da diagonal

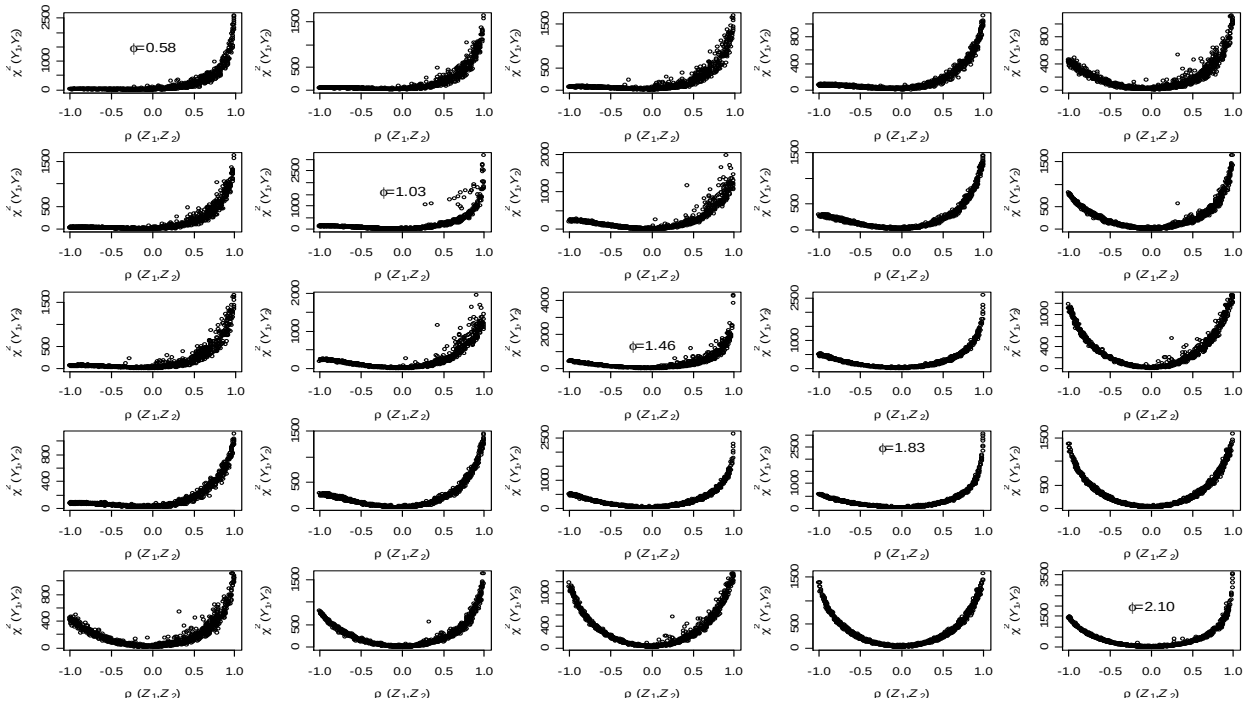

Figura 6 – Gráficos de dispersão para os coeficientes de correlação de Z_1 e Z_2, variáveis originais normalmente distribuídas, e os valores da estatística χ^2 para Y_1 e Y_2, variáveis multinomiais com $r = 6$ categorias, sob cinco diferentes graus de entropia (ϕ) indicados nos gráficos da diagonal

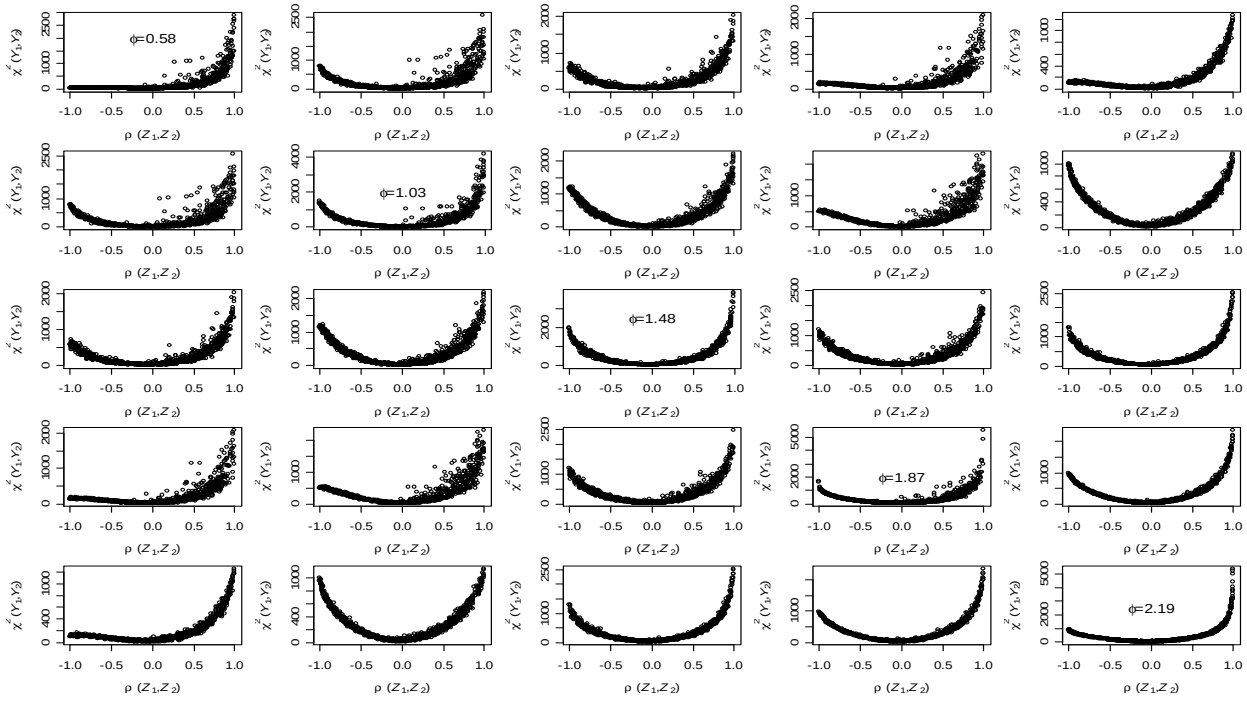

Figura 7 – Gráficos de dispersão para os coeficientes de correlação de Z_1 e Z_2, variáveis originais normalmente distribuídas, e os valores da estatística χ^2 para Y_1 e Y_2, variáveis multinomiais com $r = 8$ categorias, sob cinco diferentes graus de entropia (ϕ), indicados nos gráficos da diagonal

4.2 Árvores de classificação multivariadas baseadas em coeficientes de dissimilaridade e entropia para dados gerados de variáveis com diferentes graus de correlação e entropia

Os coeficientes de dissimilaridade e entropia apresentados no Capítulo 3 foram aplicados à construção de árvores de classificação multivariadas, e os modelos resultantes avaliados quanto às taxas de erros, às entropias e às dissimilaridades médias entre observações de um mesmo nó, calculados para as árvores das seqüências aninhadas produzidas. As Figuras 8 a 11 apresentam os resultados do estudo por simulação delineado na Seção 3.2.5, representados por meio das curvas de custo-complexidade produzidas. A influência das correlações entre variáveis respostas pode ser avaliada comparando as curvas de um mesmo gráfico e a influência de suas entropias comparando gráficos dispostos lado a lado.

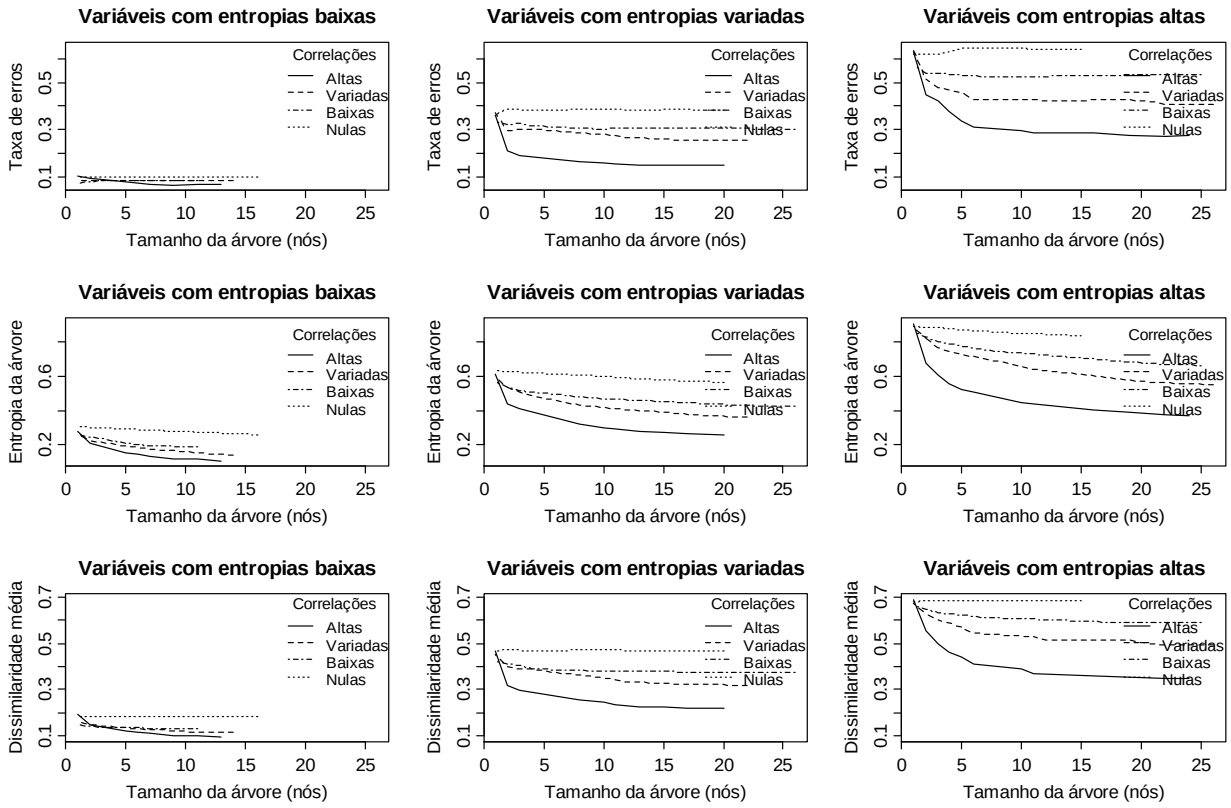

Figura 8 – Curvas de custo-complexidade para as taxas de erros (gráficos localizados acima), entropias (ao meio) e dissimilaridades médias (abaixo) de árvores de classificação multivariadas construídas com base no coeficiente de dissimilaridade simples, para dados gerados com diferentes graus de correlação e entropia

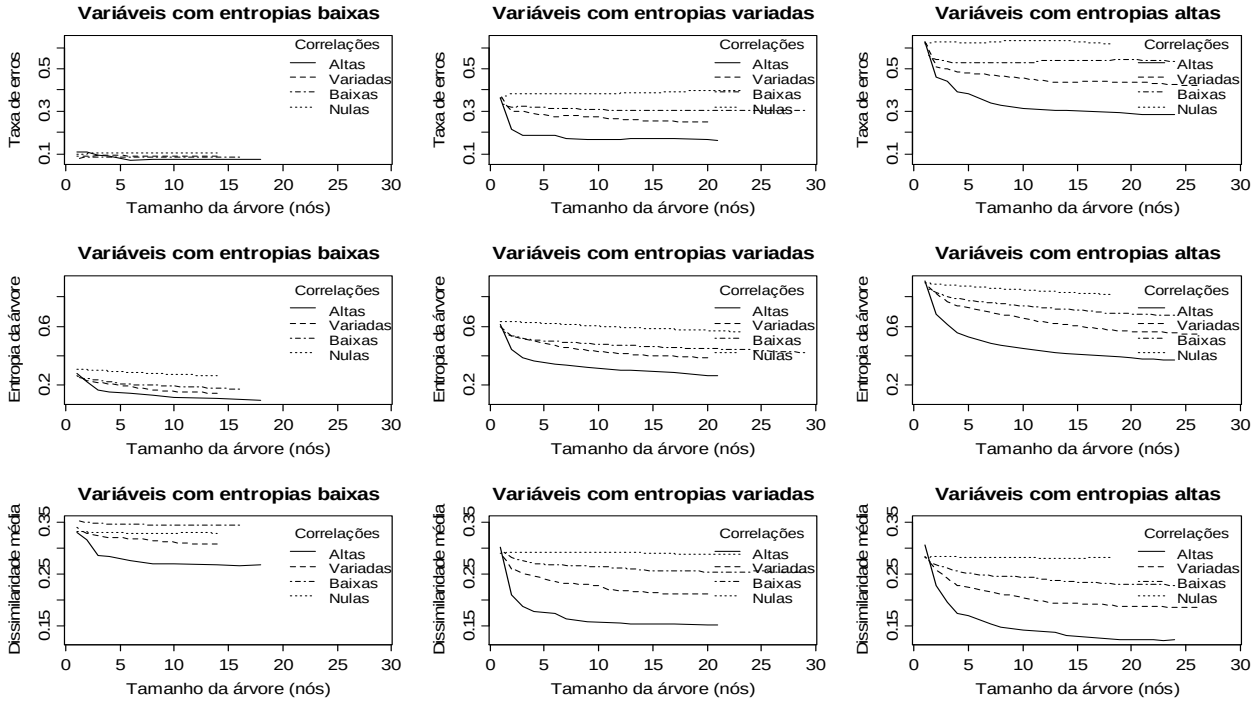

Figura 9 – Curvas de custo-complexidade para as taxas de erros (gráficos acima), entropias (ao meio) e dissimilaridades médias (abaixo) de árvores de classificação multivariadas construídas com o coeficiente de dissimilaridade baseado em probabilidades, para dados gerados com diferentes graus de correlação e entropia

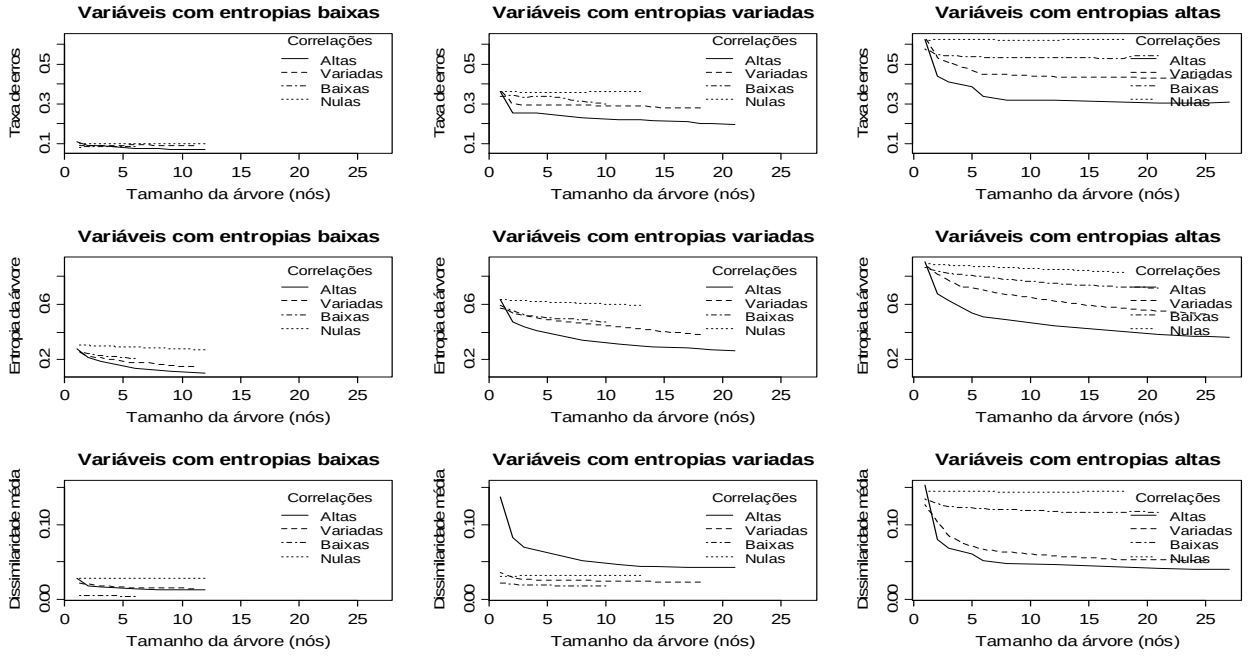

Figura 10 – Curvas de custo-complexidade para as taxas de erros (gráficos acima), entropias (ao meio) e dissimilaridades médias (abaixo) de árvores de classificação multivariadas construídas com o coeficiente de dissimilaridade baseado em distribuições condicionais de probabilidades, para dados gerados com diferentes graus de correlação e entropia

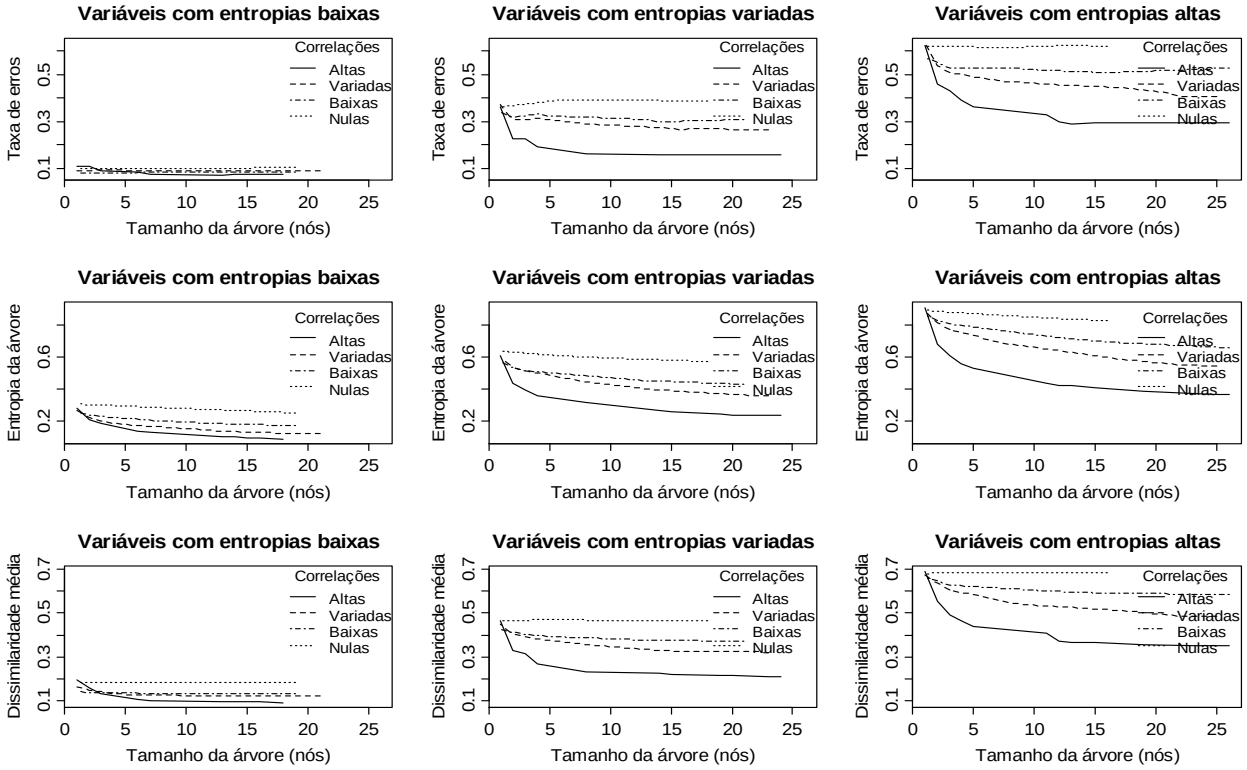

Figura 11 – Curvas de custo-complexidade para as taxas de erros (gráficos acima), entropias (ao meio) e dissimilaridades médias (abaixo) de modelos de árvores de classificação multivariadas construídas com base no coeficiente de entropia, para dados gerados com diferentes graus de correlação e entropia

Os gráficos de custo-complexidade produzidos mediante aplicação dos três coeficientes de dissimilaridade e do coeficiente de entropia evidenciam que as taxas de erros, as entropias e as dissimilaridades médias proporcionadas pelos modelos multivariados de classificação por árvores diminuem conforme aumentam as correlações entre variáveis. Além disso, quanto maiores as entropias das variáveis respostas, maiores os ganhos decorrentes da utilização das árvores multivariadas, refletidos em maiores reduções das medidas de custo consideradas.

As conclusões apresentadas, extraídas dos gráficos de custo-complexidade, foram também avaliadas por meio de análise de variância (STORCK et al., 2006), considerando, como respostas, as taxas de erros e entropias dos modelos, e como causas de variação os coeficientes aplicados, as correlações e as entropias das variáveis, além das interações de segunda ordem. O efeito da interação tripla não é considerado por não se dispor de número suficiente de graus de liberdade para testá-lo, uma vez que o estudo por simulação não foi replicado, pela complexidade de interpretação dessa interação e, sobretudo, pelo fato de análises exploratórias preliminares não

indicarem a significância do referido efeito. A principal motivação desta etapa da análise é comparar os desempenhos dos coeficientes de dissimilaridade e entropia na construção dos modelos. As pressuposições paramétricas inerentes à execução de análise de variância foram investigadas, e em nenhuma das análises verificou-se indícios de que qualquer uma delas estivesse comprometida.

A Tabela 4 descreve os resultados da análise de variância para as taxas de erros dos modelos, indicando efeito significativo, ao nível de 5% de significância, do coeficiente aplicado, das correlações e entropias das variáveis, além da interação entre entropias e correlações, na taxa média de erros dos modelos produzidos.

Tabela 4 – Análise de variância para as taxas de erros de árvores de classificação multivariadas construídas com diferentes coeficientes de dissimilaridade e entropia, para dados simulados com diferentes graus de correlação e entropia

Fonte de variação	GL	SQ	QM	Valor F	Valor p
Coeficiente	3	0,0007	0,0002	5,0911	0,0100
Correlação	3	0,2277	0,0759	1669,72	<0,0001
Entropia	2	1,1465	0,5732	12609,87	<0,0001
Coeficiente x Correlação	9	0,0007	0,0001	1,5541	0,2034
Entropia x Correlação	6	0,1040	0,0173	381,24	<0,0001
Coeficiente x Entropia	6	0,0006	0,0001	2,2000	0,0910
Resíduos	18	0,0008	0,00005		
Total	47	1,4810			

Procedeu-se, então, com a aplicação do teste de comparação múltiplas de Tukey (STORCK et al., 2006), a fim de comparar as taxas médias de erros produzidas pelos quatro coeficientes, apresentadas na Tabela 5. Conclui-se, segundo os resultados obtidos, que o coeficiente baseado em distribuições de probabilidades condicionais proporciona maiores taxas médias de erros em relação às obtidas utilizando os outros dois coeficientes de dissimilaridade. Isso se deve a uma menor eficiência do referido coeficiente ao lidar com variáveis com correlações baixas, embora o efeito da interação entre coeficiente e correlação não seja significativo ao nível de 5% de significância.

Tabela 5 – Taxas médias de erros para árvores construídas com diferentes coeficientes de dissimilaridade e entropia e resultados do teste de Tukey ao nível de significância de 5%. Caracteres diferentes indicam médias diferentes

Coeficiente	Taxa média de erro
Coeficiente de dissimilaridade simples	0,277 a
Coeficiente dissimilaridade baseado em probabilidades	0,277 a
Coeficiente dissimilaridade baseado em distribuições condicionais	0,287 b
Coeficiente de entropia	0,280 a b

A Tabela 6 descreve resultados de análise de variância para as entropias das árvores de classificação multivariadas obtidas via simulação. Com base em seus resultados, pode-se afirmar, ao nível de significância de 5%, que há efeito da interação entre os coeficientes aplicados e as correlações das variáveis nas entropias médias dos modelos gerados.

Tabela 6 - Análise de variância para as entropias de árvores de classificação multivariadas construídas com diferentes coeficientes de dissimilaridade e entropia, para dados simulados com diferentes graus de correlação e entropia

Fonte de variação	GL	SQ	QM	Valor F	Valor p
Coeficiente	3	0,0023	0,0008	5,0595	0,0102
Correlação	3	0,8411	0,2804	1860,45	<0,0001
Entropia	2	1,6148	0,8074	5357,80	<0,0001
Coeficiente x Correlação	9	0,0045	0,0005	3,3582	0,0138
Entropia x Correlação	6	0,1179	0,0196	130,43	<0,0001
Coeficiente x Entropia	6	0,0010	0,0002	1,1620	0,3687
Resíduos	18	0,0027	0,0001		
Total	47	2,5843			

A interação Coeficiente x Correlação foi analisada por meio do teste de comparações múltiplas de Tukey. Os resultados obtidos, expostos na Tabela 7, apontam que as árvores construídas com o coeficiente de dissimilaridade baseado em distribuições de probabilidades condicionais apresentam maior entropia, em média, do que as obtidas com os demais coeficientes, quando as correlações entre variáveis são baixas. Isso é coerente, uma vez que sendo

as variáveis fracamente correlacionadas, as distribuições condicionais são pouco informativas. Para variáveis com correlações altas, variadas ou nulas as entropias médias não diferem entre os coeficientes.

Tabela 7 – Entropias médias de árvores de classificação multivariadas obtidas com a aplicação de diferentes coeficientes com diferentes correlações entre variáveis e resultados do teste de Tukey ao nível de significância de 5%. Caracteres diferentes indicam médias diferentes. Os caracteres maiúsculos referem-se às comparações efetuadas entre correlações para cada coeficiente, e os minúsculos às comparações efetuadas entre coeficientes para cada grau de correlação

Coeficiente	Correlações			
	Altas	Variadas	Baixa	Nulas
Coeficiente de dissimilaridade simples	0,255 A a	0,362 B a	0,461 C a	0,604 D a
Coeficiente dissimilaridade baseado em probabilidades	0,262 A a	0,373 B a	0,469 C a	0,604 D a
Coeficiente dissimilaridade baseado em distribuições condicionais de probabiliades	0,303 A a	0,370 B a	0,501 C b	0,613 D a
Coeficiente de entropia	0,241 A a	0,355 B a	0,462 C a	0,613 D a

A análise de variância para as dissimilaridades médias dos modelos não foi executada pelo fato dos coeficientes abordados serem expressos em escalas distintas, inviabilizando a comparação de suas magnitudes.

4.2.1 Comparação dos procedimentos uni e multivariados na classificação dos nós finais em árvores multivariadas

Como ressaltado ao longo do presente estudo, um dos objetivos associados à construção de modelos de regressão e classificação por árvores é a predição de novos elementos com base nos resultados das covariáveis. Para tanto, deve-se alocar cada uma destes novos elementos a um dos nós finais do modelo originado, e classificá-lo segundo algum critério estabelecido a partir dos elementos presentes no nó. Na Seção 3.2.2, foram apresentadas duas possibilidades de classificação dos nós: segundo os resultados mais freqüentes avaliados para cada variável individualmente ou conjuntamente. Ambos os procedimentos foram utilizados em cada uma das

simulações realizadas. Os resultados foram absolutamente idênticos em aproximadamente 80% dos casos, indicando que, para as configurações consideradas, os dois procedimentos produziram resultados bastante semelhantes. Estudos mais detalhados, no entanto, são recomendados, de modo a avaliar o desempenho dos dois procedimentos de classificação ao analisar diferentes números de variáveis com números distintos de categorias.

4.2.2 Avaliação do critério do ponto mais afastado na seleção de árvores de classificação multivariadas

Outra proposta apresentada neste trabalho é um critério alternativo aplicado à seleção do melhor modelo de classificação por árvores, descrito na seção 3.2.4, ao qual se denominou “regra do ponto mais afastado”. Foi proposta a escolha daquele modelo responsável por minimizar conjuntamente o custo e a complexidade associados às arvores de classificação multivariadas. A regra do desvio padrão (BREIMAN et al., 1984) busca a seleção do menor modelo cujo custo de má-classificação não difira substancialmente do modelo com menor custo, o que pode ocasionar a seleção de modelos com elevados números de nós finais. A análise apresentada na seqüência tem por objetivo comparar os desempenhos dos dois critérios na seleção da melhor árvore de classificação.

A Figura 12 apresenta os tamanhos, as entropias e as taxas de erros das árvores obtidas mediante aplicação dos dois procedimentos de seleção, para cada um dos 36 modelos simulados, resultantes das análises de variáveis com diferentes graus de correlação e entropia e da utilização de coeficientes de dissimilaridade e entropia distintos, conforme descrito na Seção 3.2.5. Foram desconsiderados os resultados relativos às 12 árvores geradas sob correlações nulas, situação em que a utilização dos modelos propostos mostrou-se improdutiva. Nos três gráficos dessa figura, pontos acima da reta tracejada indicam valores maiores para a regra do ponto mais afastado em relação à regra do desvio padrão. Pontos abaixo da reta tracejada indicam o contrário. É evidente a tendência do procedimento baseado no ponto mais afastado em selecionar árvores com menos nós finais. A regra do ponto mais afastado, se comparada à regra do desvio padrão, proporcionou uma redução de aproximadamente 55% no tamanho médio das árvores, ocasionando, entretanto, um aumento de aproximadamente 6% nas entropias médias das árvores e de 18% em suas taxas médias de erros.

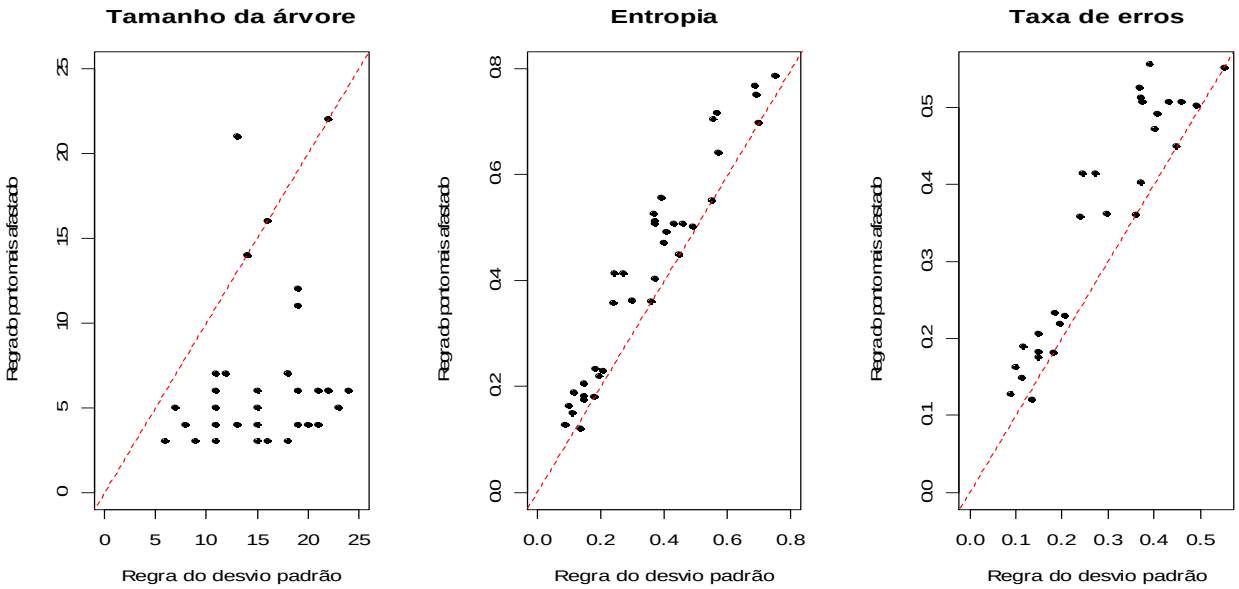

Figura 12 – Gráficos de dispersão para os tamanhos, as taxas de erros e as entropias de árvores de classificação multivariadas selecionadas pelas regras do desvio padrão e do ponto mais afastado para variáveis com diferentes graus de correlação e entropia

Não se tem por objetivo propor a substituição do critério do desvio padrão pelo do ponto mais afastado. Chama-se atenção, no entanto, para o fato de que o método do ponto mais afastado, além de requerer menor esforço computacional, pondera de maneira mais equilibrada as medidas de custo e complexidade do que a regra do desvio padrão, que tende a selecionar modelos com baixo custo e elevada complexidade. Sugere-se que as duas regras sejam consideradas no processo de escolha do modelo, conjugadas na busca por uma árvore com interpretação coerente e elucidativa para o fenômeno sob estudo.

4.3 Árvores de classificação multivariadas aplicadas ao estudo do consumo de álcool e fumo dentre os habitantes do município de Botucatu (SP).

Foram construídas árvores de classificação multivariadas para os dados de fumo e alcoolismo, com base nos três coeficientes de dissimilaridade para dados categorizados apresentados e na medida multivariada de entropia. A regra do desvio padrão foi utilizada na busca pelo modelo mais parcimonioso. Fique registrado, no entanto, que apesar de não ter seus resultados apresentados, a regra do ponto mais afastado também foi considerada, produzindo

árvores menores (entre três e seis nós finais a menos) do que as produzidas pela regra do desvio padrão. Optou-se por considerar a regra do desvio padrão por se julgar adequados os tamanhos de árvores produzidas (entre 8 e 11 nós finais). De maneira semelhante ao estabelecido no estudo com dados simulados, decidiu-se não partir nós com 20 elementos ou menos, nem formar nós com menos de 10 elementos. Os resultados obtidos são apresentados nas seções seguintes.

4.3.1 Árvore de classificação multivariada para os dados de consumo de álcool e fumo, obtida com os coeficientes de dissimilaridade simples e baseado em probabilidades.

Os coeficientes de dissimilaridade simples e baseado em probabilidades produziram duas seqüências de árvores aninhadas bastante semelhantes, além de conduzir à seleção da mesma árvore. A Figura 13 apresenta as curvas de custo-complexidade geradas utilizando cada um desses coeficientes.

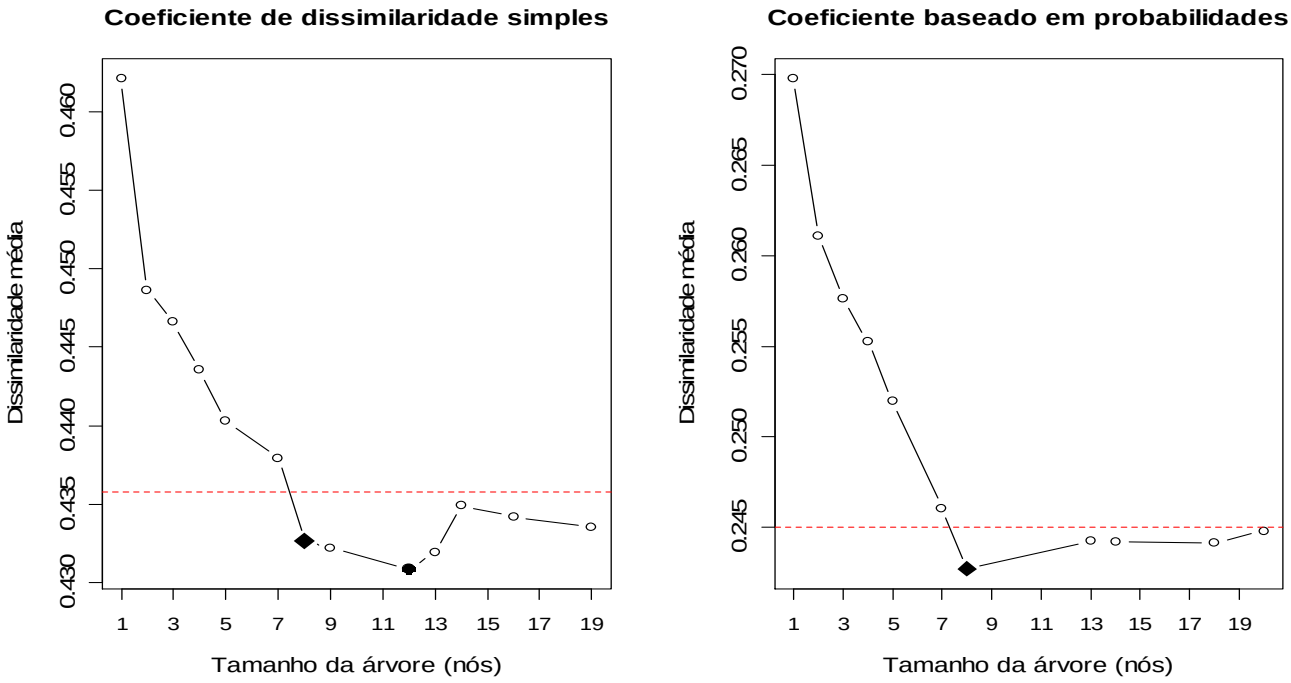

Figura 13 – Gráfico de custo-complexidade para as árvores de classificação multivariadas construídas para os dados de consumo de álcool e fumo, com base nos coeficientes de dissimilaridade simples e baseado em probabilidades. O ponto representado por (●) indica a árvore com menor dissimilaridade média, o ponto representado por (♦) 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

A árvore selecionada sob ambos os coeficientes é apresentada na Figura 14. Visando facilitar a compreensão da mesma, explica-se, por exemplo, que a partição do nó 2 em dois novos

nós (4 e 5) ocasionou uma redução na dissimilaridade média da árvore de $\Delta_{\phi}(Nó\ 2) = 0{,}011$. Além disso, também a título de exemplo, os elementos do nó 2 que afirmaram ter no máximo uma pessoa para compartilhar seus problemas (NNPI < 2) foram alocados ao nó 4, enquanto aqueles com duas pessoas ou mais para compartilhar seus problemas (NNPI ≥ 2) foram alocados ao nó 5.

Três variáveis são responsáveis por duas partições cada, indicando associação das mesmas com o conjunto de variáveis respostas. São elas: o grau máximo de escolaridade (SEDU), o número de pessoas com quem o entrevistado pode compartilhar seus problemas (NNPI) e a ocupação profissional (WPOS). Além delas, somente o ano de nascimento (DATE) compõe a árvore, produzindo uma única partição. A forma como essas variáveis estão associadas às de consumo de álcool e fumo é explorada na seqüência. A Figura 15 apresenta gráficos relativos à composição dos nós finais da árvore apresentadas na Figura 14, quanto às cinco variáveis respostas consideradas. São apresentadas as freqüências relativas ao número de elementos em cada nó, facilitando a comparação dos referidos nós. A análise conjunta desses gráficos com a árvore de classificação constituída é determinante para a identificação das associações de interesse.

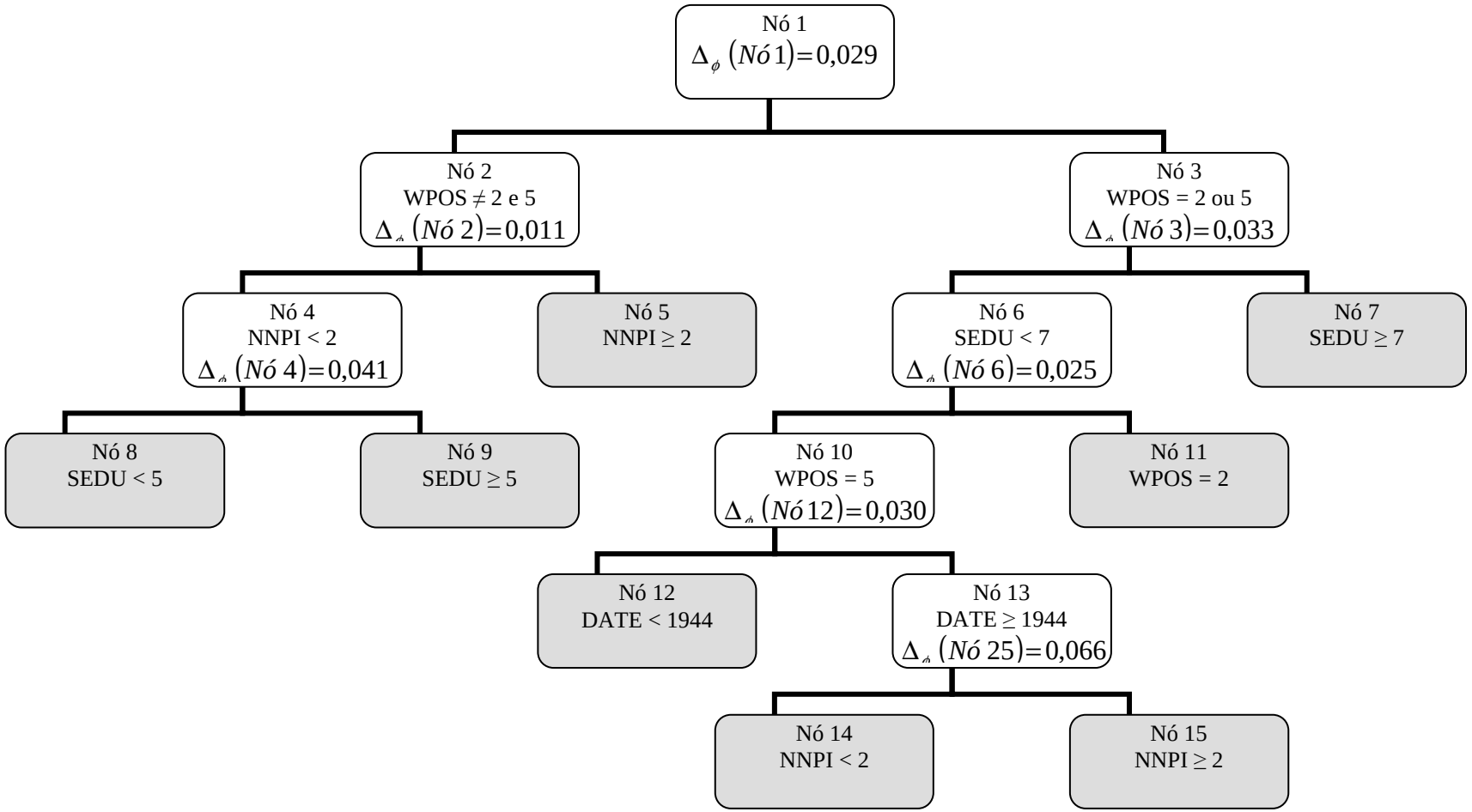

Figura 14 – Árvore de classificação multivariada obtida com os coeficientes de dissimilaridades simples e baseado em probabilidades. Os valores de Δ_{ϕ} referem-se às reduções das dissimilaridades médias produzidas pelas partições, calculadas segundo o coeficiente de dissimilaridade simples. Os códigos utilizados para as variáveis que compõem o modelo são os seguintes: WPOS: ocupação profissional (2 - dona de casa; 4 – afastado por motivo de doença, 5 – aposentado, 6 – estudante, 7 – desempregado, 8 – empregado); NNPI NNPI: Sem contar o parceiro conjugal, quantas pessoas têm para compartilhar seus problemas (1 – Nenhuma, 2 – Uma,3 – 2 a 3, 4 – 4 a 5, 5 – 6 ou mais); 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); DATE: ano de nascimento. No interior de cada nó são representadas as partições executadas e as conseqüentes reduções na dissimilaridade média do modelo. Os nós com preenchimento são nós finais

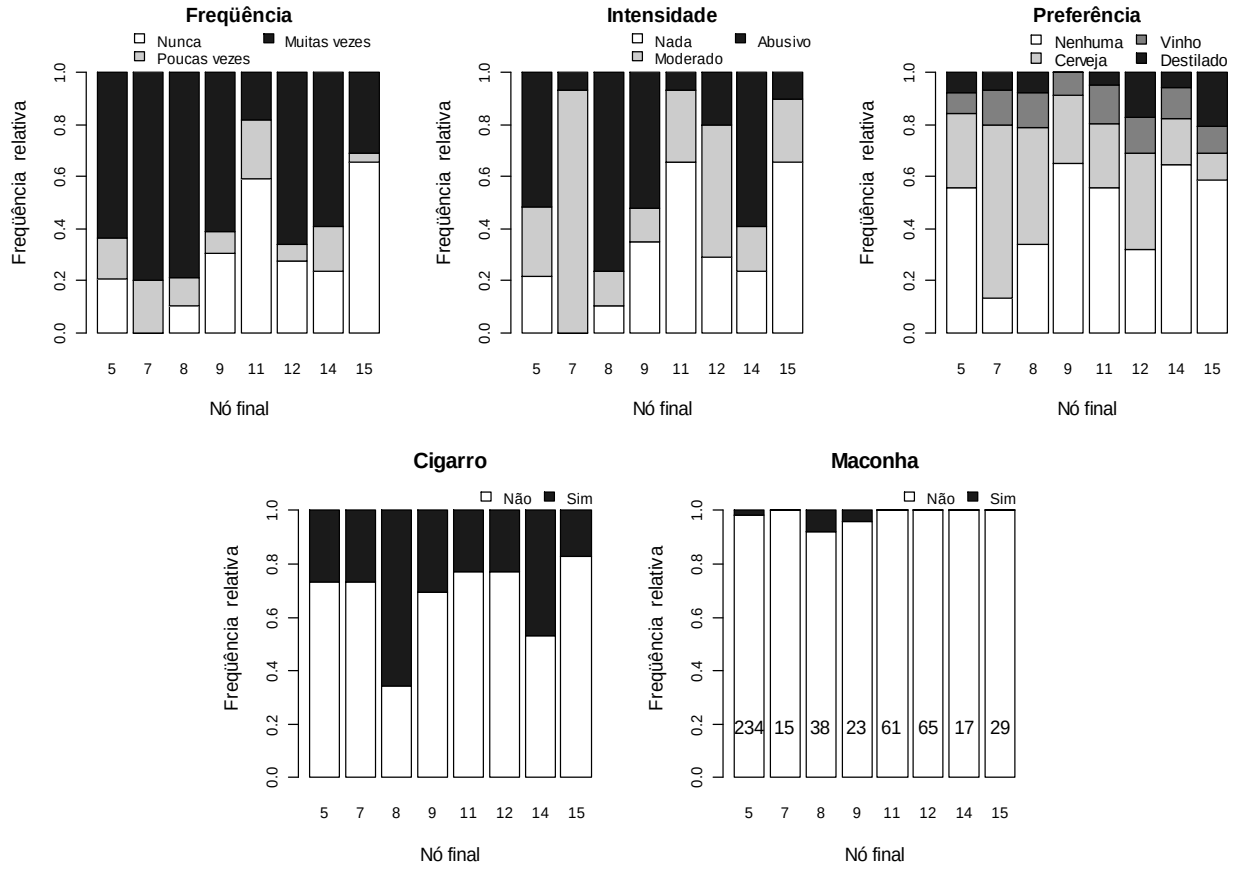

Figura 15 – Composição dos nós finais da árvore de classificação multivariada obtida com os coeficientes de dissimilaridade simples e baseado em probabilidades, quanto à freqüência e à intensidade de consumo alcoólico no último ano, à bebida preferida e aos consumos de cigarro e maconha. Os valores no interior das colunas do gráfico relativo ao consumo de maconha indicam os tamanhos dos nós

Pode-se destacar, pela análise conjunta dos gráficos dispostos na Figuras 15 e da árvore de classificação multivariada apresentada na Figura 14, que o nó 8 é composto, em sua maioria, por indivíduos que beberam muitas vezes ao longo do último ano (79%), afirmaram ter bebido de forma abusiva pelo menos em uma ocasião (76%) e apresenta as maior proporções de fumantes (66%) e usuários de maconha (8%) dentre os oito nós formados. Os indivíduos que compõem este nó não são aposentados e nem donas de casa, têm baixa escolaridade (no máximo completaram o primeiro grau) e afirmaram não ter nenhuma pessoa com quem compartilhar seus problemas.

Os indivíduos do nó 9, por sua vez, têm características semelhantes aos do nó 8, mas com percentuais de consumo de álcool e fumo menores. Diferem dos indivíduos do nó 8 apenas pelo fato de terem pelo menos uma pessoa com quem compartilhar seus problemas. Todos os indivíduos que compõem o nó 7 afirmaram ter bebido ao longo do último ano. A grande maioria

(93%), no entanto, não bebeu de forma abusiva sequer uma vez. Dentre todos os nós constituídos, é aquele com maior porcentagem de apreciadores de cerveja (67%). Os indivíduos que compõem este nó são aposentados ou donas de casa com elevada escolaridade (todos, no mínimo, deram início a um curso superior).

Os nós 11 e 15 são compostos por indivíduos que, em sua maioria, afirmaram não ter consumido bebidas alcoólicas no último ano (60 e 65%, respectivamente). O nó 11 é composto por donas de casa e o nó 15 por aposentados com elevada escolaridade, nascidos após 1944 e que contam com pelo menos uma pessoa para confidenciar seus problemas. Indivíduos do nó 14 diferem daqueles que compõem o nó 15 por não terem com quem dividir seus problemas. As porcentagens associadas aos consumos de álcool e cigarro são todas maiores para indivíduos do nó 14, em relação àqueles que compõem o nó 15.

Visando ratificar as conclusões mencionadas, procedeu-se à execução de uma análise de correspondência múltipla (Seção 2.6), compreendendo as cinco variáveis dependentes e uma nova, indicadora dos nós aos quais os indivíduos foram alocados. Como dito anteriormente, proximidades no gráfico de análise de correspondência indicam associação entre as categorias. Pode-se, com base na dispersão dos pontos no gráfico da análise de correspondência, apresentado na Figura 16, levantar evidências a respeito de associação entre o nó 8 e consumo de cigarro e álcool de forma abusiva e freqüente, o nó 7 e o consumo moderado e pouco freqüente de álcool e os nós 11 e 15 e o não consumo de álcool e fumo, dentre outras. O gráfico da análise de correspondência confirma os comentários anteriormente citados.

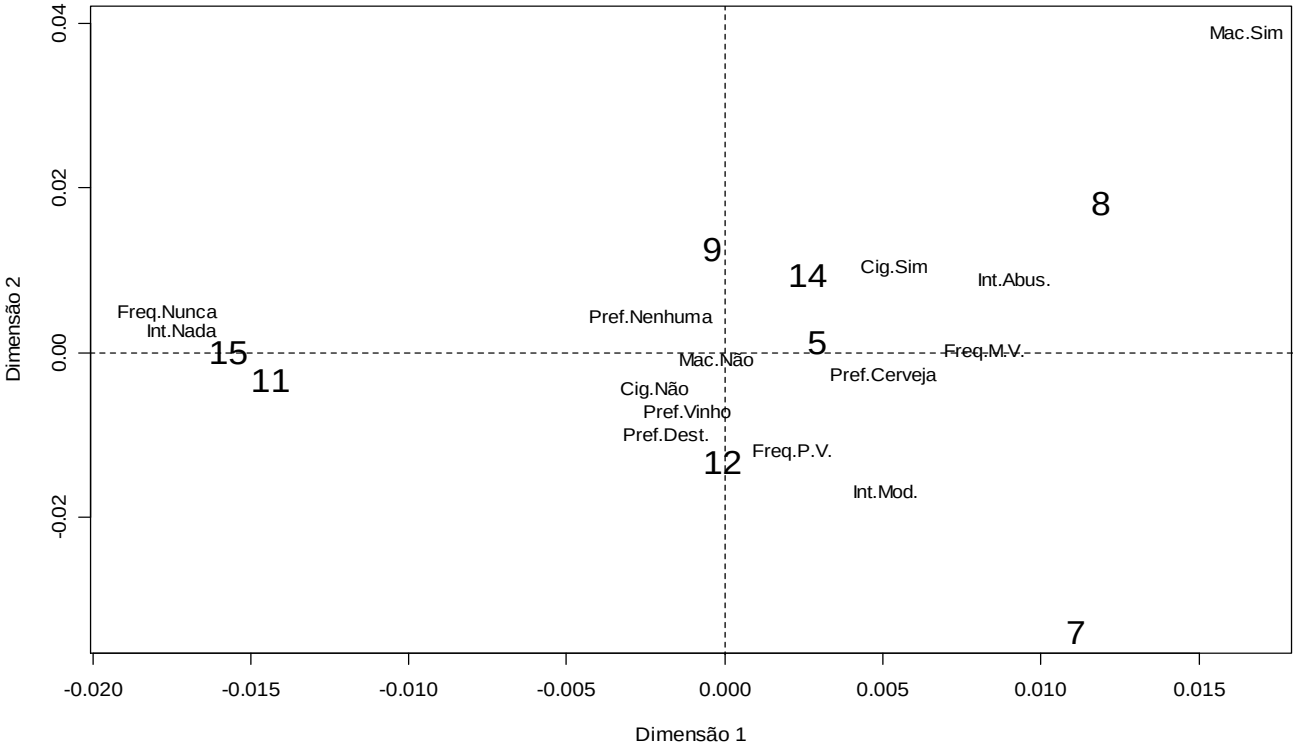

Figura 16 – Gráfico da análise de correspondência múltipla para as variáveis ‘Freqüência com que bebeu no último ano’ – Freq. (sendo M.V. = muitas vezes e P.V. = poucas vezes), ‘Intensidade com que bebeu quando mais consumiu álcool’ – Int. (sendo Mod. = Moderado e Abus. = abusivamente), ‘Consumo de cigarro’ – Cig, ‘Consumo de maconha’ – Mac. e ‘Bebida preferida’ – Pref. (sendo Dest. = destilado). Os números representados no interior do gráfico indicam os nós finais

Como ressaltado na Seção 4.2.6, as classificações dos nós finais de árvores de classificação multivariadas segundo os resultados mais freqüentes avaliados para cada variável individualmente ou conjuntamente foram bastante semelhantes. Optou-se, então, por classificar os nós finais conforme o resultado mais freqüente para cada variável individualmente. Essa opção deve-se à insegurança quanto a classificação dos nós com base em freqüências conjuntas ao considerar números mais elevados de variáveis e categorias por variável. Estima-se, via validação cruzada, que a taxa de predições corretas fornecidas pelo modelo seja de 67%. A Figura 17 permite avaliar de maneira mais detalhada o poder preditivo da árvore.

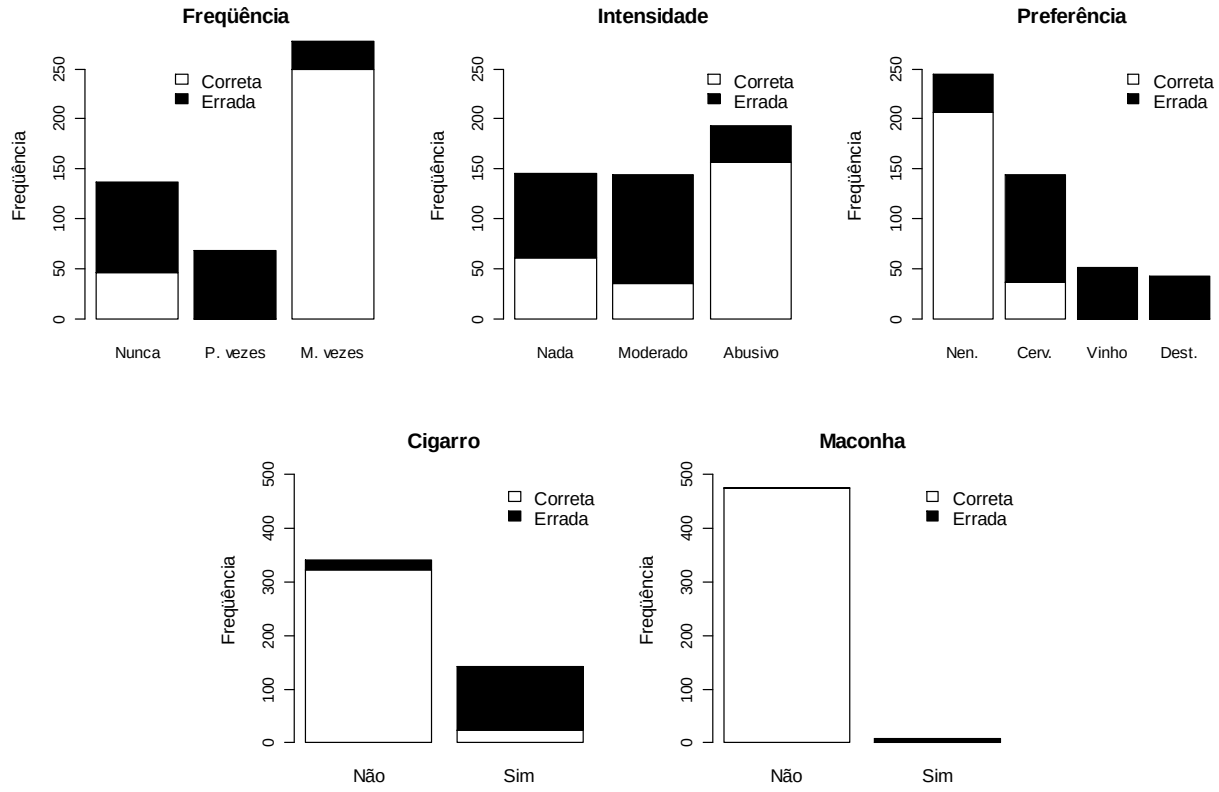

Figura 17 – Taxas de predições corretas e erradas para a freqüência (sendo P. vezes = poucas vezes, M. vezes = muitas vezes) e intensidade de consumo alcoólico no último ano, bebida preferida (sendo Nen. = nenhuma, Cerv. = cerveja e Dest. = destilado) e consumo de cigarro e maconha, produzidas pela árvore de classificação multivariada construída com os coeficientes de dissimilaridade simples e baseado em probabilidades

Observam-se maiores taxas de predições corretas associadas às categorias mais freqüentes. Em contrapartida, categorias menos freqüentes apresentam, invariavelmente, taxas de predições corretas baixas ou até mesmo nulas. Isso já era esperado, uma vez que a classificação dos nós finais é realizada a partir dos resultados mais freqüentes dentre os elementos que os constituem. Dependendo do objetivo da análise, uma alternativa frequentemente aplicada em modelos univariados, com o objetivo de aumentar a taxa de predições corretas de determinadas categorias, é a incorporação de custos de más-classificações (BREIMAN et al., 1984). Para fins exploratórios, no entanto, mais importante do que simplesmente atribuir uma classificação a um novo elemento é avaliar a composição dos elementos presentes no nó ao qual ele é alocado.

4.3.2 Árvore de classificação multivariada para os dados de consumo de álcool, cigarro e maconha fundamentada no coeficiente de dissimilaridade baseado em distribuições condicionais de probabilidades.

A Figura 18 apresenta o gráfico de custo-complexidade para a árvore obtida utilizando o coeficiente de dissimilaridade baseado em distribuições condicionais de probabilidades. Segundo a regra do desvio padrão, a árvore com nove nós finais é selecionada, sendo representada na Figura 19.

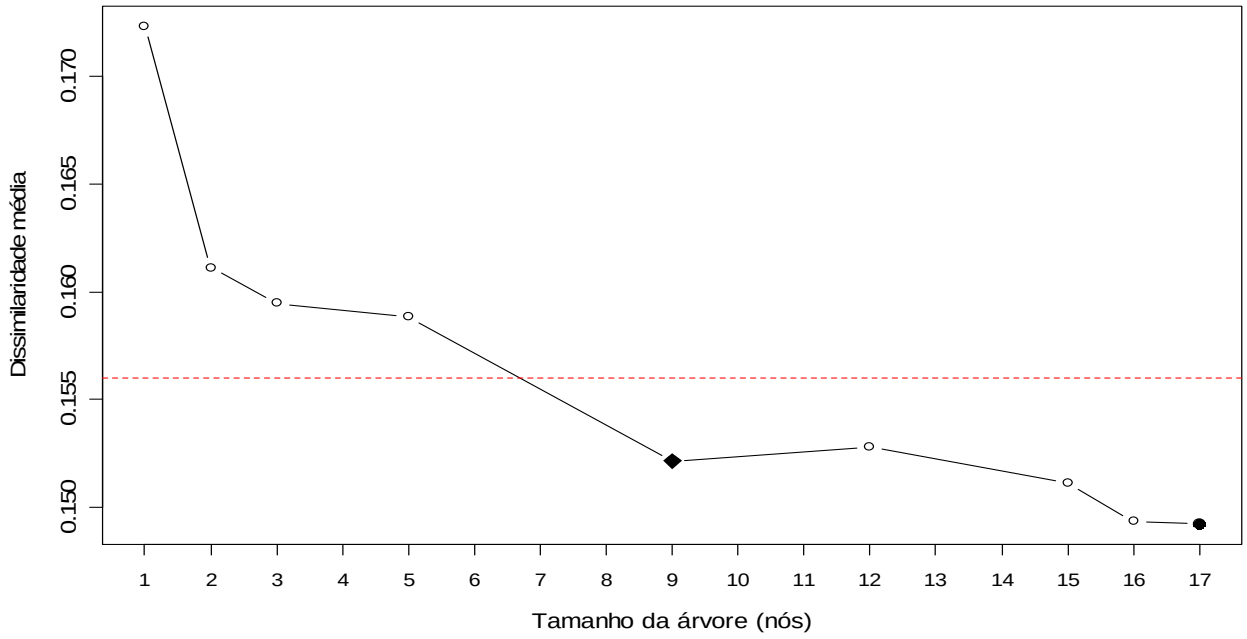

Figura 18– Gráfico de custo-complexidade para a árvore de classificação multivariada construída para os dados de consumo de álcool e fumo, com base no coeficiente de dissimilaridade calculado a partir de distribuições condicionais de probabilidades. O ponto representado por (●) indica a árvore com menor 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

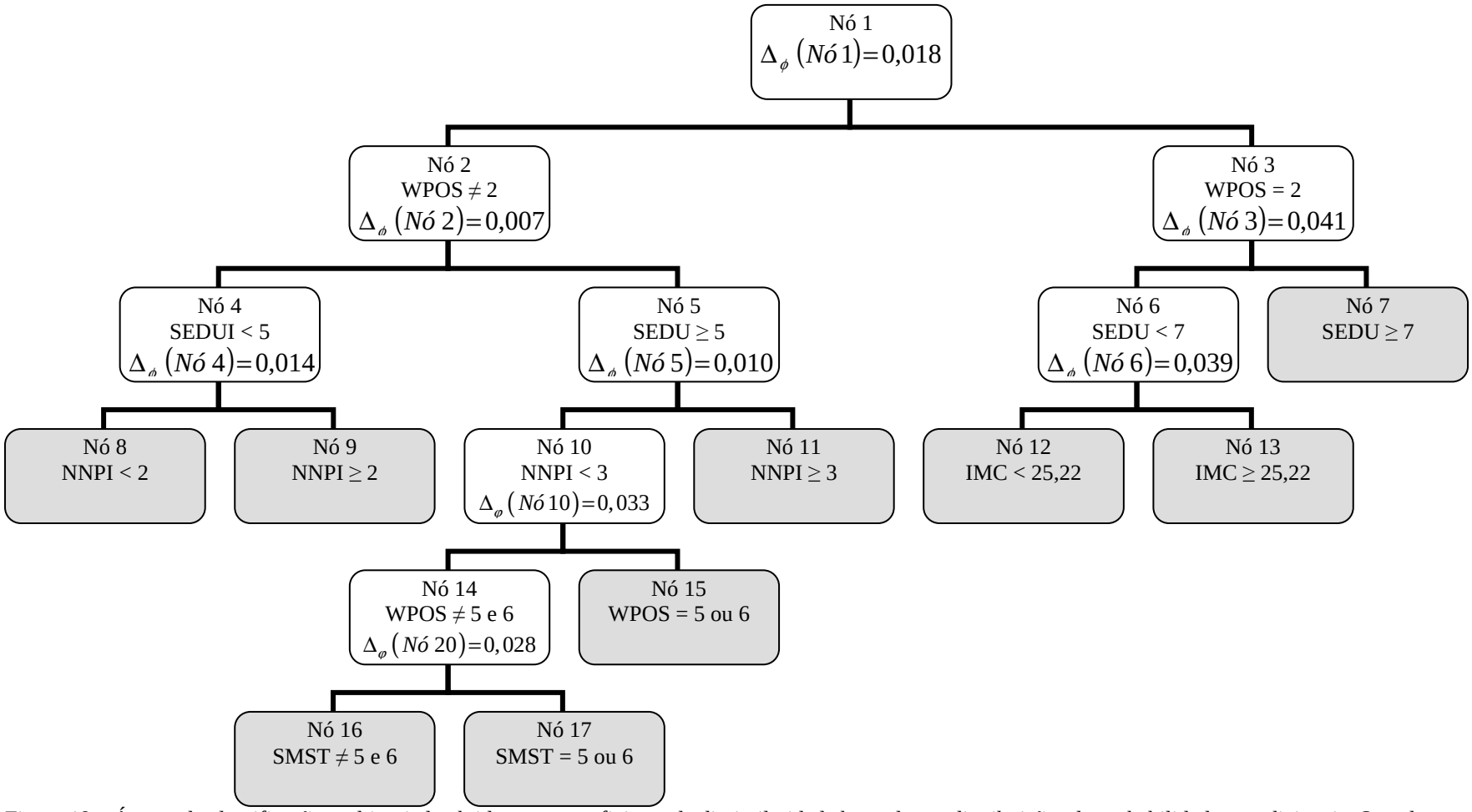

Figura 19 – Árvore de classificação multivariada obtida com o coeficiente de dissimilaridade baseado em distribuições de probabilidades condicionais. Os valores de Δ_ϕ referem-se às reduções das dissimilaridades médias produzidas pelas partições. Os códigos utilizados para as variáveis que compõem o modelo são os seguintes: WPOS: Ocupação profissional (2 - dona de casa; 4 – afastado por motivo de doença, 5 – aposentado, 6 – estudante, 7 – desempregado, 8 – empregado); 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); NNPI: Sem contar o parceiro conjugal, quantas pessoas têm para compartilhar seus problemas (1 – Nenhuma, 2 – Uma,3 – 2 a 3, 4 – 4 a 5, 5 – 6 ou mais); IMC: Índice de massa corporal; SMST: Situação conjugal;. No interior de cada nó são representadas as partições executadas e as conseqüentes reduções na dissimilaridade média do modelo. Os nós com preenchimento são nós finais

Novamente, a interpretação do modelo obtido requer a avaliação dos nós finais quanto às freqüências observadas para as variáveis correspondentes aos consumos de álcool e fumo. A Figura 20 representa as freqüências relativas às variáveis respostas em cada nó final.

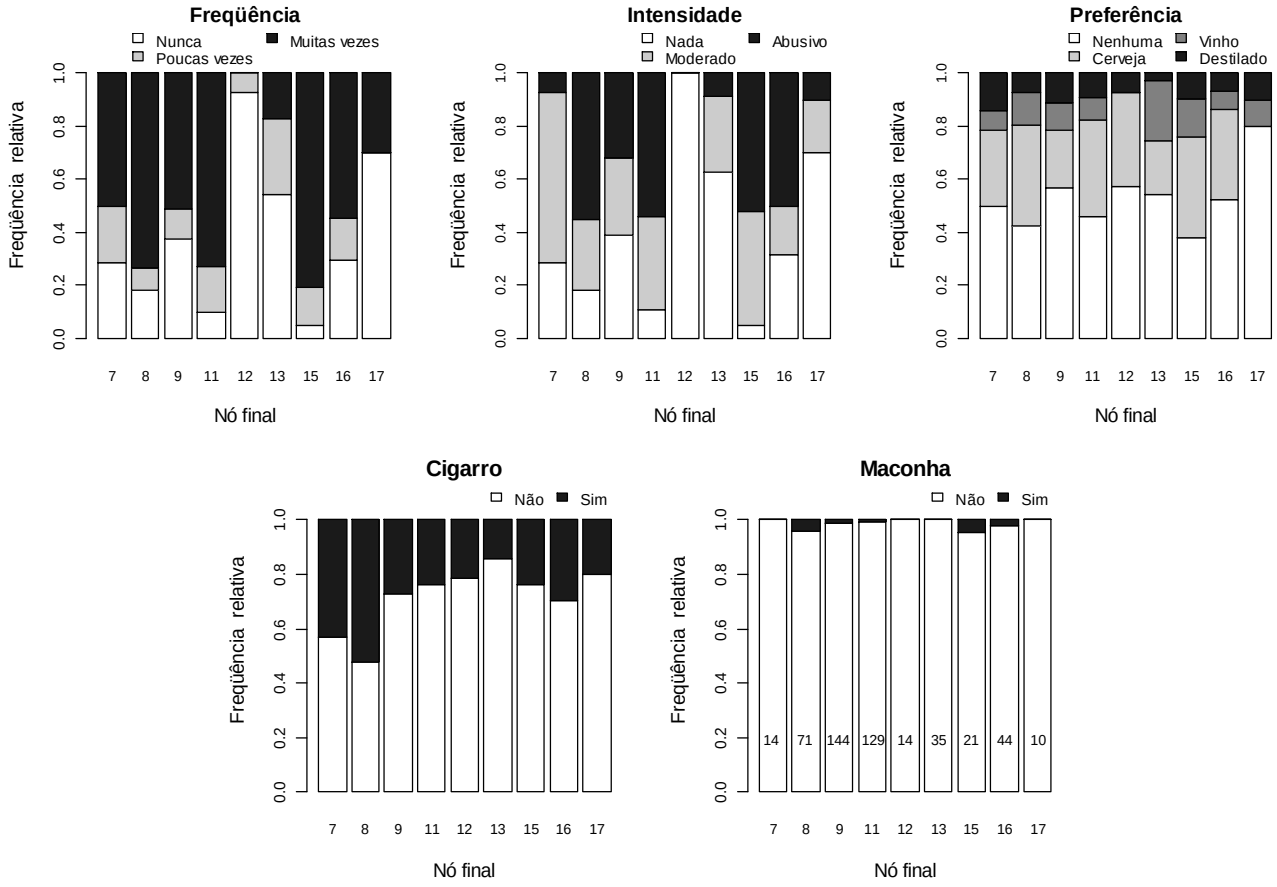

Figura 20 – Composição dos nós finais da árvore de classificação multivariada baseada no coeficiente de dissimilaridade calculado a partir de distribuições condicionais de probabilidades quanto à freqüência e à intensidade de consumo alcoólico no último ano, à bebida preferida e aos consumos de cigarro e maconha. Os valores no interior das colunas do gráfico relativo ao consumo de maconha indicam os tamanhos dos nós

Conclui-se, com base na árvore de classificação multivariada apresentada na Figura 19 e nos gráficos de colunas apresentados na Figura 20, 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 as maiores porcentagens 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.

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, têm-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 21 apresenta o gráfico produzido por uma análise de correspondência múltipla, executada de maneira semelhante à descrita para o coeficiente de dissimilaridade simples.

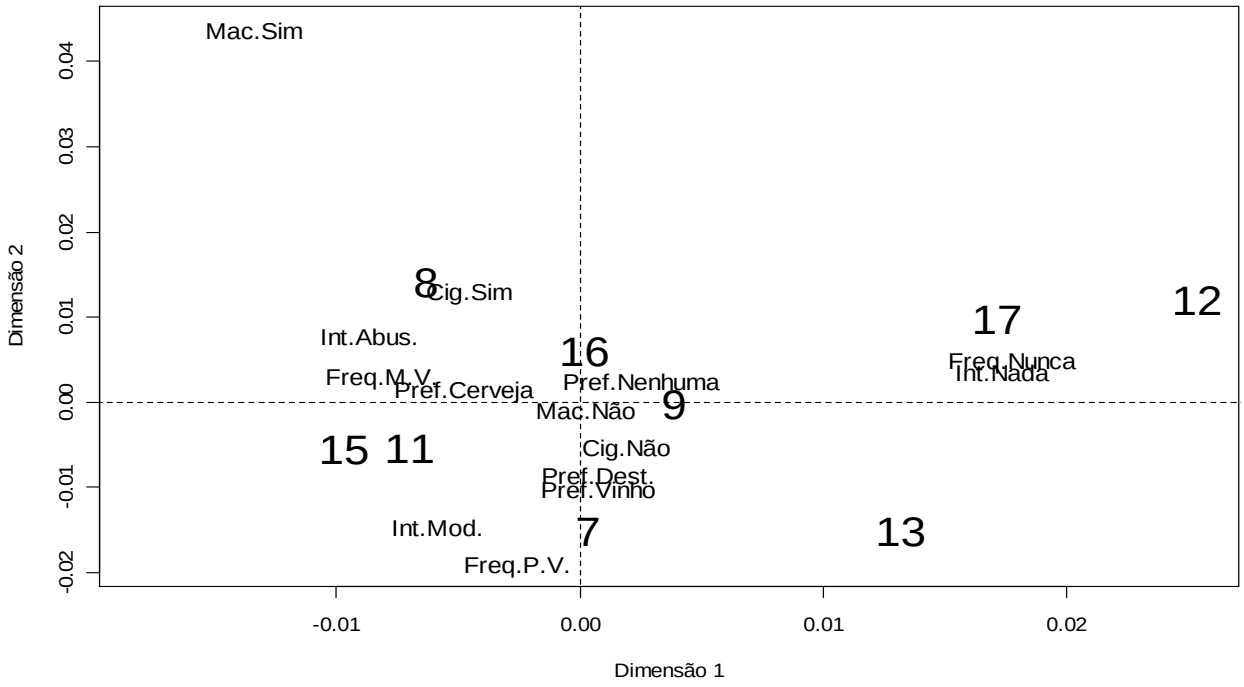

Figura 21 – Gráfico da análise de correspondência múltipla para as variáveis ‘Freqüência com que bebeu no último ano’ – Freq (sendo M.V. = muitas vezes e P.V. = poucas vezes), ‘Intensidade com que bebeu quando mais consumiu álcool’ – Int. (sendo Mod. = Moderado e Abus. = abusivamente), ‘Consumo de cigarro’ – Cig, ‘Consumo de maconha’ – Mac. e ‘Bebida preferida’ – Pref. (sendo Dest. = destilado). Os números representados no interior do gráfico indicam os nós finais

A proximidade dos nós 12, 13 e 17 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 associações levantadas entre os nós e as variáveis citadas anteriormente. A Figuras 22 apresenta as taxas de predições corretas produzidas pelo modelo, estimadas por validação cruzada. Novamente, maiores taxas de predições corretas são verificadas para resultados mais freqüentes.

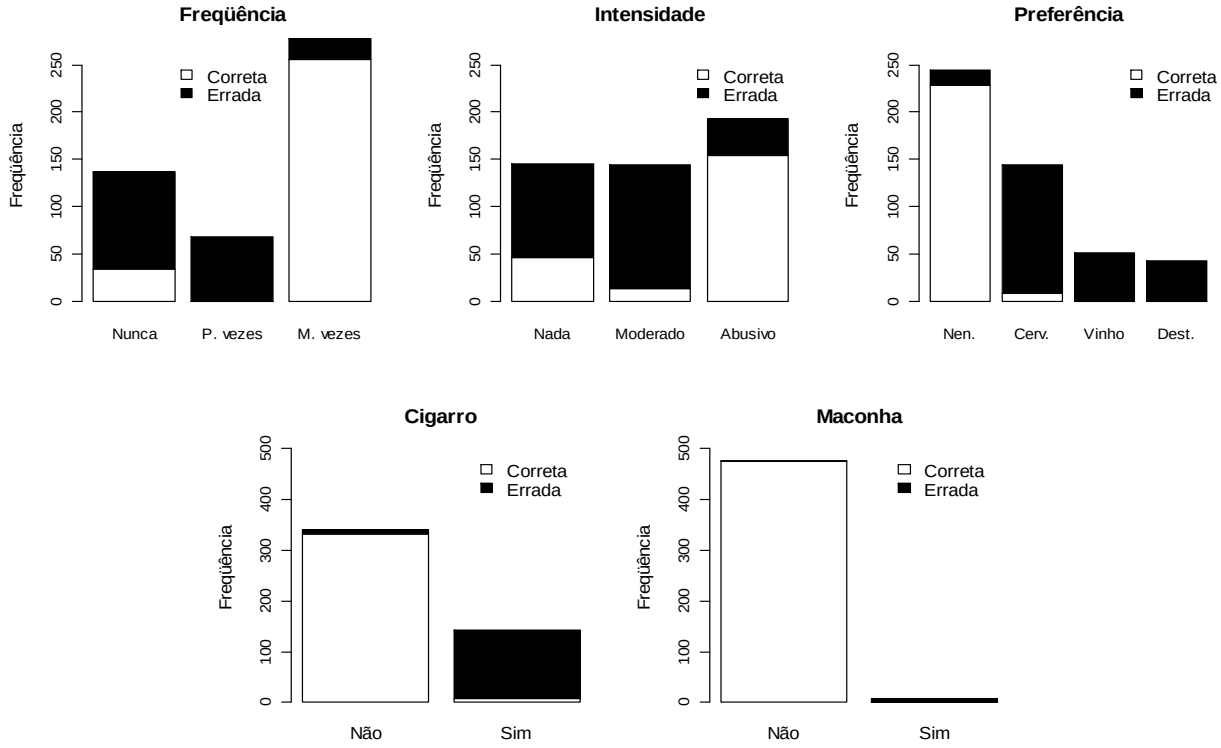

Figura 22 – Taxas de predições corretas e erradas para a freqüência (sendo P. vezes = poucas vezes, M. vezes = muitas vezes) e intensidade de consumo alcoólico no último ano, bebida preferida (sendo Nen. = nenhuma, Cerv. = cerveja e Dest. = destilado) e consumo de cigarro e maconha, produzidas pela árvore de classificação multivariada construída com o coeficiente de dissimilaridade baseado em distribuições de probabilidades condicionais

4.3.3 Árvore de classificação multivariada para os dados de consumo de álcool, cigarro e maconha fundamentada no coeficiente de entropia.

A Figura 23 apresenta a curva de custo complexidade da seqüência de árvores aninhadas obtidas com a aplicação da medida de entropia. Com base na regra do desvio padrão, opta-se pela árvore com 11 nós finais. A referida árvore é representada na Figura 24.

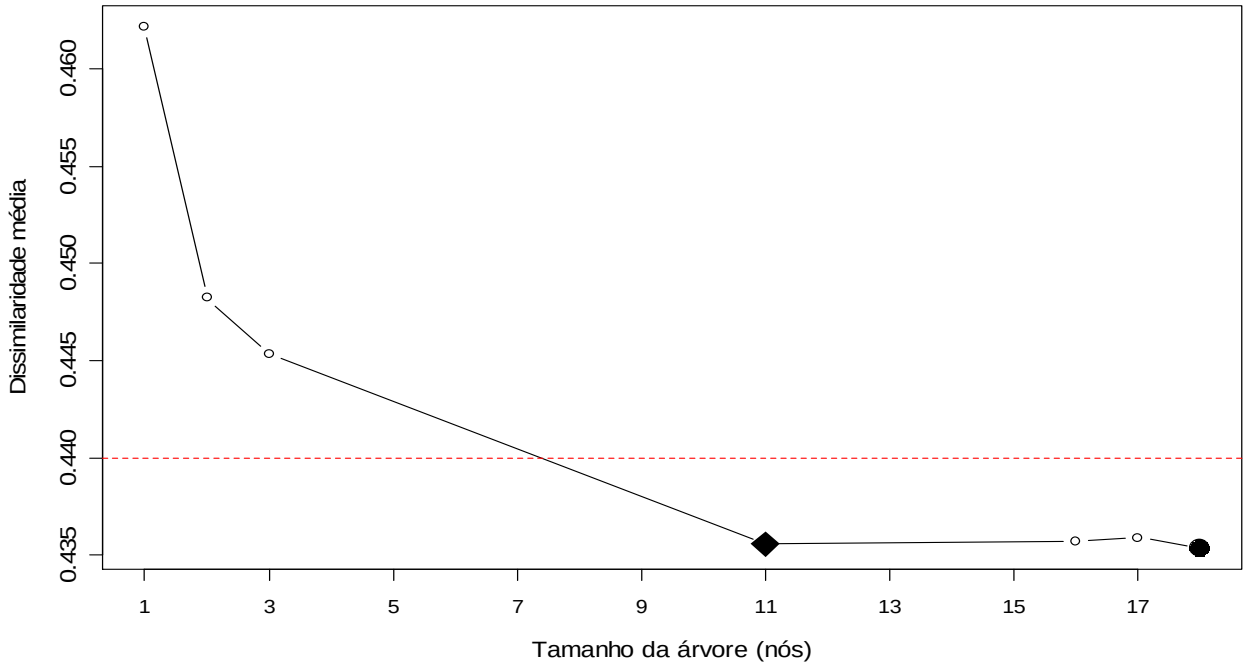

Figura 23– Gráfico de custo-complexidade para a árvore de classificação multivariada construída para os dados de consumo de álcool e fumo, com base na medida de entropia. O ponto representado por (●) indica a árvore com menor dissimilaridade média, o ponto representado por (♦) indica a árvore selecionada pela regra do desvio padrão e a linha horizontal tracejada (---) o limite superior da dissimilaridade média associado à regra do desvio padrão

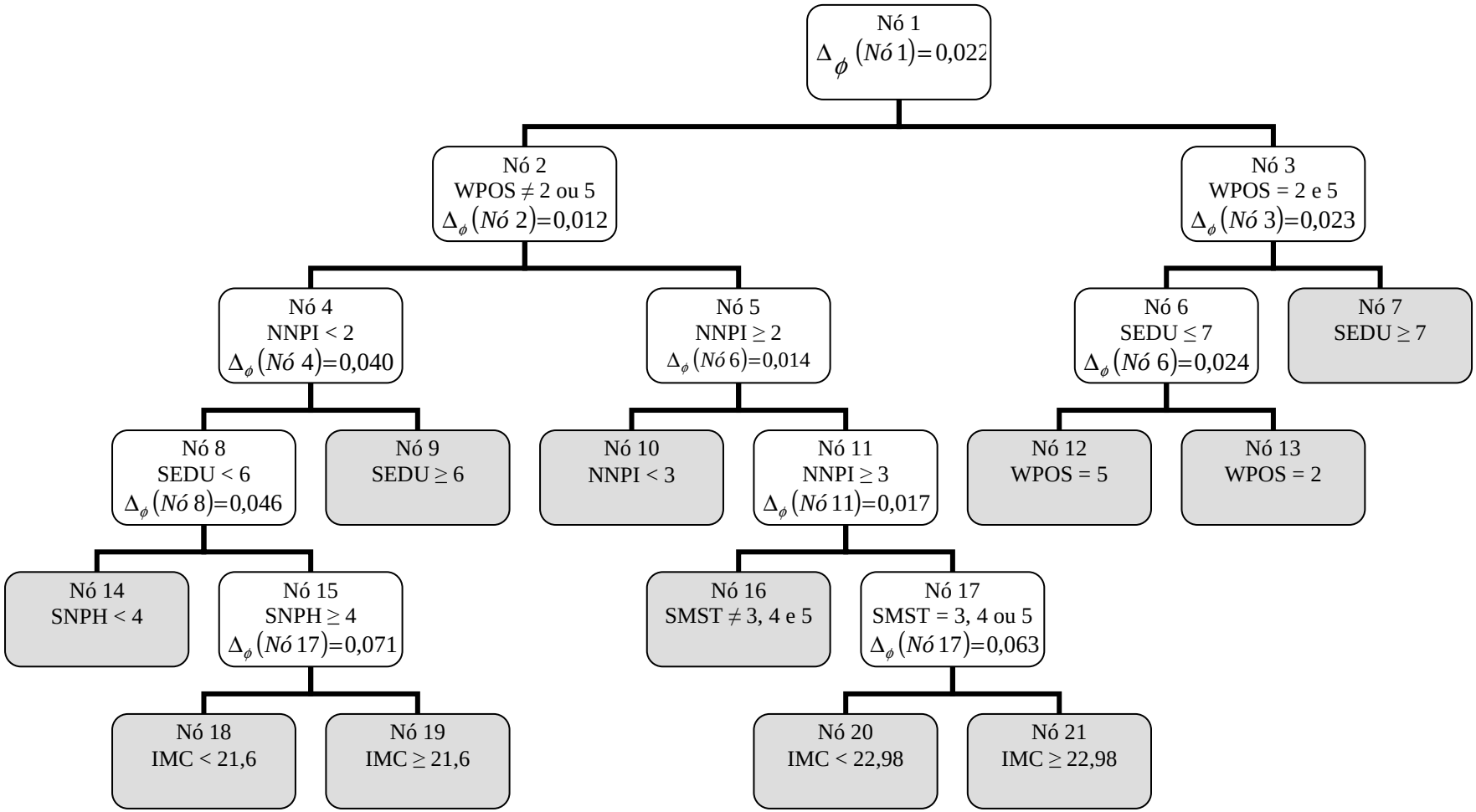

Figura 24 – Árvore de classificação multivariada obtida com o coeficiente de entropia. Os valores de Δ_ϕ referem-se às reduções das dissimilaridades médias produzidas pelas partições, calculadas segundo o coeficiente de dissimilaridade simples. Os códigos utilizados para as variáveis que compõem o modelo são os seguintes: WPOS: ocupação profissional (2 - dona de casa; 4 – afastado por motivo de doença, 5 – aposentado, 6 – estudante, 7 – desempregado, 8 – empregado); NNPI: Sem contar o parceiro conjugal, quantas pessoas têm para compartilhar seus problemas (1 – Nenhuma, 2 – Uma,3 – 2 a 3, 4 – 4 a 5, 5 – 6 ou mais); 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); SNPH: número de pessoas que residem com o entrevistado; SMST: situação conjugal; IMC: índice de massa corporal. No interior de cada nó são representadas as partições executadas e as conseqüentes reduções na dissimilaridade média do modelo. Os nós com preenchimento são nós finais

A Figura 25 apresenta o gráfico produzido por uma análise de correspondência múltipla, realizada de maneira semelhante à descrita anteriormente, quando considerados os coeficientes de dissimilaridade.

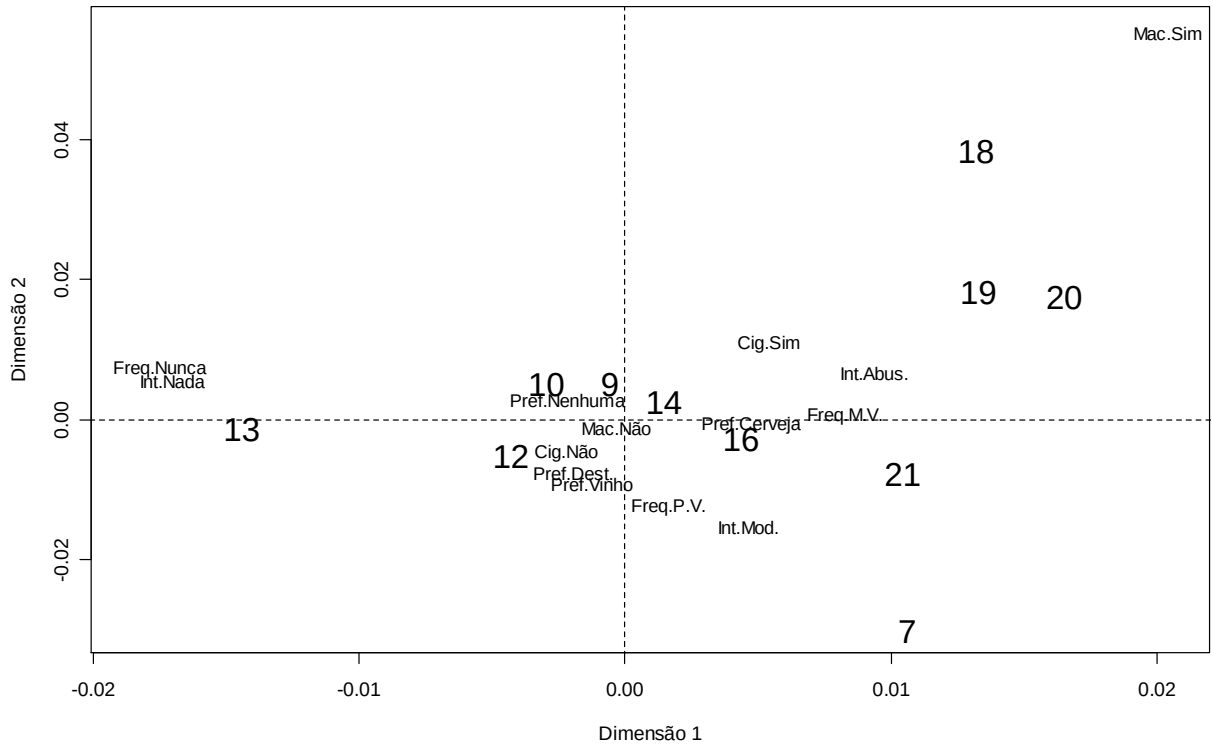

Figura 25 – Gráfico da análise de correspondência múltipla para as variáveis ‘Freqüência com que bebeu no último ano’ – Freq (sendo M.V. = muitas vezes e P.V. = poucas vezes), ‘Intensidade com que bebeu quando mais consumiu álcool’ – Int. (sendo Mod. = Moderado e Abus. = abusivamente), ‘Consumo de cigarro’ – Cig, ‘Consumo de maconha’ – Mac. e ‘Bebida preferida’ – Pref. (sendo Dest. = destilado). Os números representados no interior do gráfico indicam os nós finais

O consumo de cigarro, maconha, consumo freqüente e abusivo de álcool têm suas representações no mesmo quadrante do gráfico da análise de correspondência, 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 nós citados 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 casadas, 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 colunas apresentados na Figura 26 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.

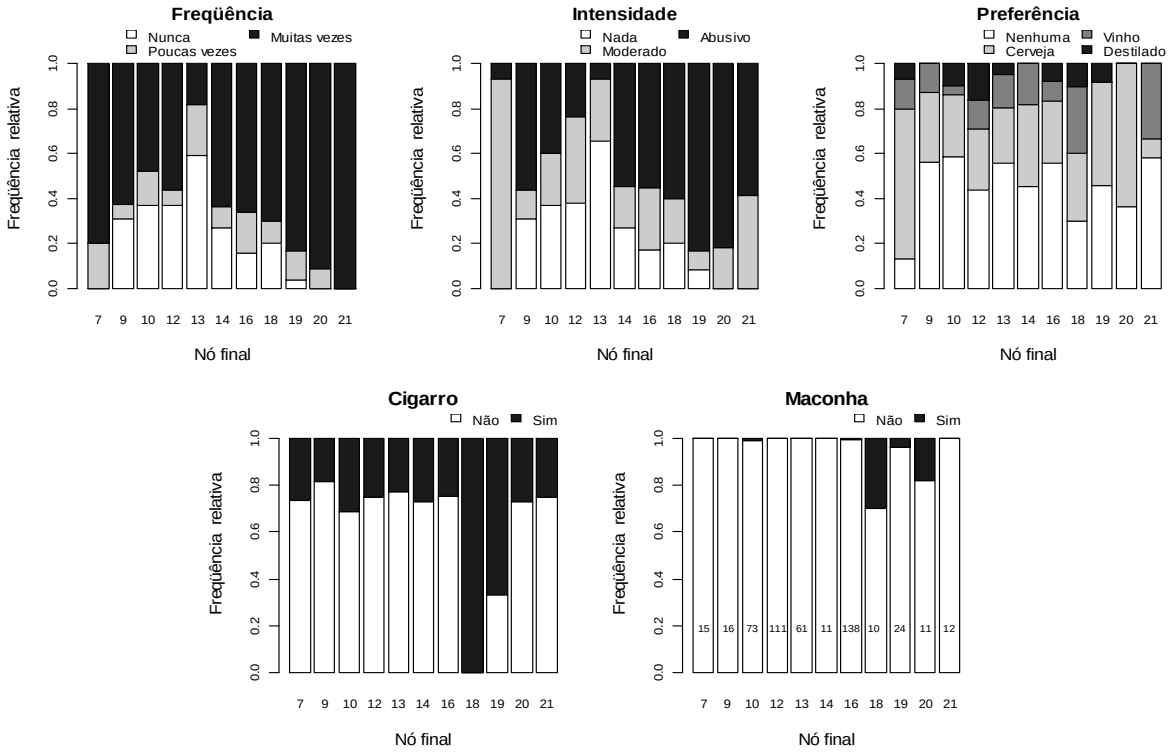

Figura 26 – Composição dos nós finais da árvore de classificação multivariada baseada na medida de entropia, quanto à freqüência e à intensidade de consumo alcoólico no último ano, à bebida preferida e aos consumos de cigarro e maconha. Os valores no interior das colunas do gráfico relativo ao consumo de maconha indicam os tamanhos dos nós

As taxas de predições corretas fornecidas pela árvore de classificação multivariada baseada na medida de entropia podem ser verificadas na Figura 27. Também aqui as categorias mais freqüentes apresentam maiores índices de predições corretas.

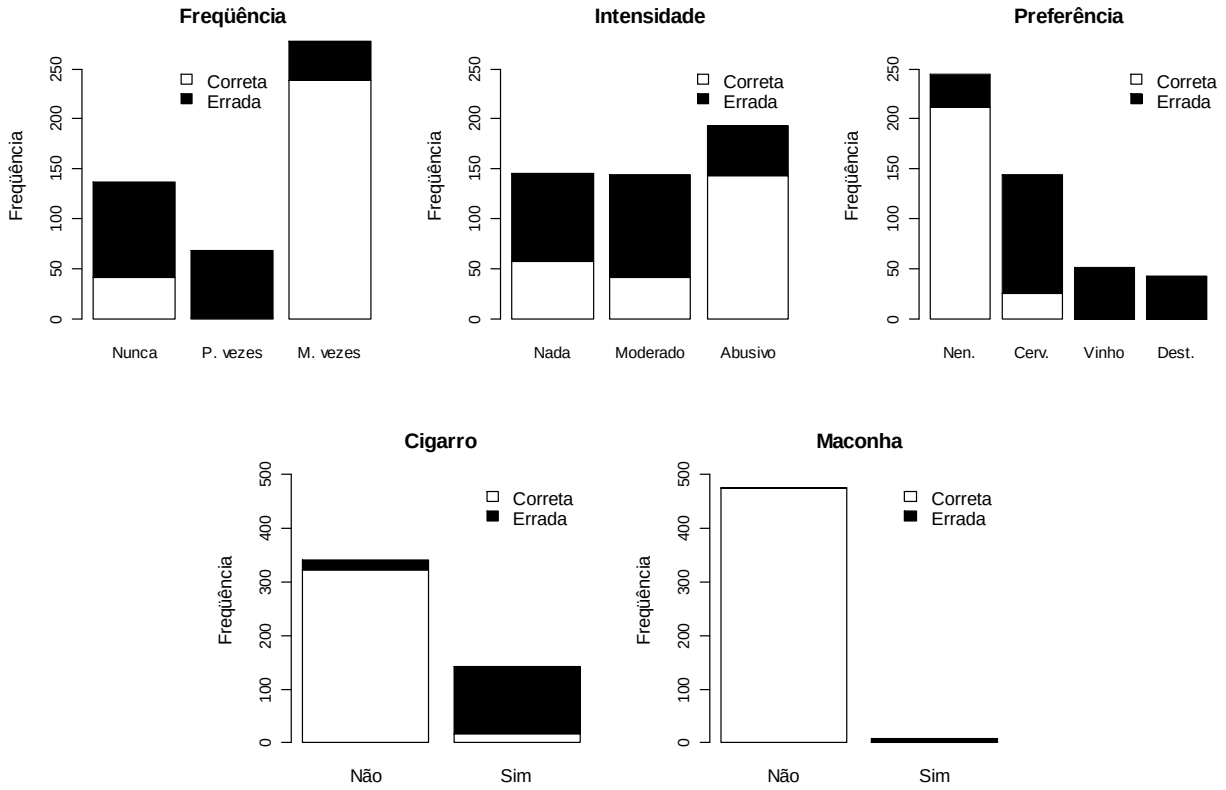

Figura 27 – Taxas de predições corretas e erradas para a freqüência e intensidade de consumo alcoólico no último ano (sendo P. vezes = poucas vezes, M. vezes = muitas vezes, Nen. = nenhuma, Cerv. = cerveja e Dest. = destilado), bebida preferida e consumo de cigarro e maconha produzidas pela árvore de classificação multivariada construída com base na medida de entropia

5 CONSIDERAÇÕES FINAIS

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

Pôde-se verificar, por meio dos resultados do estudo por simulação, que os métodos multivariados propostos são capazes de explicar de maneira adequada a variação original dos dados, sendo que os resultados produzidos são melhores, em termos de menores dissimilaridades, entropias e taxas de predições erradas, quão maiores as entropias e correlações das variáveis respostas. Diferenças nos resultados entre os diferentes coeficientes de dissimilaridades e entropia utilizados para quantificar a heterogeneidade dos nós somente foram verificadas sob baixas correlações, situação em que as árvores geradas pelo coeficiente de dissimilaridade baseado em distribuições condicionais apresentaram maiores entropia e taxa de predições incorretas do que as demais.

A análise dos dados de alcoolismo e fumo permitiu detectar perfis diferentes de indivíduos, quanto às suas características pessoais, sociais e econômicas, dentre outras, que se associam a padrões distintos de consumo de álcool e fumo. Para todos os modelos construídos, baseados nos três coeficientes de dissimilaridades e no de entropia, as variáveis ‘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 de consumo alcoólico e fumo, à medida que cada uma destas variáveis é responsável por duas partições em cada um dos modelos. As conclusões extraídas dos três modelos são compatíveis, indicando, por exemplo, a tendência de maior consumo de álcool e fumo dentre elementos com baixa escolaridade, que exercem alguma atividade profissional e não tem amigos com quem compartilhar seus problemas.

Avaliou-se, adicionalmente, um procedimento de simulação de múltiplas variáveis multinomiais, que se mostrou eficiente, em boa parte das configurações consideradas, na geração de variáveis com estruturas distintas de dependência. Propôs-se ainda um procedimento

alternativo para a seleção do modelo, denominado “regra do ponto mais afastado”, consistindo na escolha do modelo que minimiza conjuntamente as medidas de custo e complexidade. Verificou-se, com base nos resultados produzidos pelo estudo por simulação, que a regra proposta gerou uma redução média de 55% nos tamanhos das árvores, em relação à tradicional “regra do desvio padrão”. Em contrapartida, foram verificados aumentos de 6% na entropia média e 18% na taxa de predições incorretas dos modelos.

Dentre as possíveis extensões do presente trabalho, destaca-se a proposição de alternativas para se lidar com dados ausentes, técnicas adequadas à análise de dados categorizados longitudinais e à análise de dados multivariados apresentando, simultaneamente, variáveis respostas categorizadas e numéricas.

REFERÊNCIAS

AGRESTI, A. **Categorical data analysis.** New York : Wiley, 1990. 558p.

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

BUSTOS, O.H.; ORGAMBIDE, A.C.F. Simulação estocástica: teoria e algoritmos. In SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 10, 1992, Rio de Janeiro. **Anais...** São Paulo: ABE, 1992. 152 p.

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. **American Journal of Political Science**, Austin, v.24, n.1, p. 155-174, Feb.1980.

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, Nov. 2000.

DE'ATH, G. Multivariate Regression Trees: A New Technique for Modeling Species-Environment Relationships. **Ecology**, Brooklin, v.83, n.4, p.1105–1117, Apr. 2002.

DIAS, C.T.S. **Planejamento de uma fazenda em condições de risco: programação linear e simulação multidimensional.** 1996. 100p. Tese (Doutorado em Estatística e Experimentação Agronômica) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 1996.

GOWER, J.C. **Biplots.** London; New York: Chapman & Hall, 1996. 277p.

GOODAL, D.W. A New Similarity Index Based on Probability. **Biometrics,** Washington, v.22, n.4, p.882-907, Dec 1966.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis.** 4.ed. New Jersey: Prentice Hall, 1998. 816p.

KULLBACK, S.; LEIBLER, R.A. On Information and Sufficiency. **The annals of mathematical statistics**, Beachwood, v.22, n.1, p. 79-86, Mar 1951.

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

LEE, S.K. On generalized multivariate decision tree by using GEE. **Computational Statistics & Data Analysis**, Amsterdan, v.49, n.4, p.1105-1119, June 2005.

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

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

R DEVELOPMENT CORE TEAM. **R: A language and environment for statistical computing**., Vienna, Austria, 2006. Disponível em: http://www.R-project.org, 2007. Acesso em: 20 mar. 2008.

RENCHER, A.C. **Methods of multivariate analysis**. New York: John Wiley, 1995. 627p.

ROMESBURG, H.C. **Cluster Analysis for Researchers**. North Carolina: Lulu. Press, 2004. 334p.

SEGAL, M.R. Tree-structured methods for longitudinal data. **Journal of the American Statistical Association**, Boston, v.87, p.407–418, June 1992.

STORK, l.; GARCIA, D.C.; LOPES, S.J.; ESTEFANEL, V. **Experimentação vegetal.** 2.ed. Santa Maria: UFSM, 2006. 198p.

WORLD HEALTH ORGANIZATION. **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, **Journal of the American Statistical Association**, Boston, v.93, p.180-193, Mar. 1998.

APÊNDICE

APÊNDICE A – Programas computacionais

A.1 Cálculo da matriz de dissimilaridades para o coeficiente baseado em distribuições condicionais de probabilidades.

Entrada: dados - data frame contendo o conjunto de variáveis respostas;
Saída: simi - armazena a matriz de dissimilaridades.
Nota - as variáveis respostas não podem ter categorias com nomes coincidentes.

```
diss=function(dados){
vet1=numeric()
vet2=numeric()
vet3=numeric()

for(i in 1:ncol(dados)){
      for (j in 1:ncol(dados)){
            if(j!=i){
                  tab=table(dados[,i],dados[,j])
                  somas=rowSums(tab)
                  probs=tab/somas
                  probs[probs==0]=0.001
                  probs2=matrix(0,nlevels(dados[,i]),nlevels(dados[,i]))
                  for(p in 1:(nlevels(dados[,i])-1)){
                        for(q in (p+1):nlevels(dados[,i])){
                              probs2[p,q]=sum(probs[p,]*log(probs[p,]/
                              probs[q,],base=2)+probs[q,]*log(probs[q,]/
                              probs[p,],base=2))
                              vet1=c(vet1,levels(dados[,i])[p])
                              vet2=c(vet2,levels(dados[,i])[q])
                              vet3=c(vet3,probs2[p,q])
                        }
                  }
            }
      }
}
w=tapply(vet3, factor(vet1):factor(vet2), sum)
w=data.frame(w)
w=na.omit(w)
w=cbind(w,h=row.names(w))

a=matrix(0,nrow(dados),nrow(dados))
for(i in 1:(nrow(dados)-1)){
      for(j in (i+1):nrow(dados)){
            er=factor(t(dados[i,]))
            er2=factor(t(dados[j,]))
            novo=er:er2
            novo2=er2:er
            a[i,j]=sum(w[,1][which(w[,2]%in%novo|w[,2]%in%novo2)])
```

```
      }
}
diag(a)=0
list("simi"=a/max(a))}
```

A.2 – Cálculo da matriz de dissimilaridades simples

Entrada: dados - data frame contendo o conjunto de variáveis respostas;
Saída: simi - matriz de dissimilaridades.

```
simple=function(dados){
dados=t(dados)
Simple=matrix(0,ncol(dados),ncol(dados))
for (i in 1:ncol(dados)){
      for (j in 1:ncol(dados)){
            na=sum(dados[,i]==dados[,j])
            nb=sum(dados[,i]!=dados[,j])
            Simple[i,j]=(na)/(na+nb)
      }
}
result=list("simi"=1-Simple)}
```

A.3 – Cálculo da entropia de um vetor de proporções (função interna).

Entrada: dad - resultados de uma variável categorizada;
Saída: entropia do vetor.

```
ent=function(dad)
-sum((dad)*log(dad,2))/log(length(dad),2)
```

A.4 – Cálculo da entropia para os resultados de um conjunto de variáveis categorizadas (função interna).

Entrada: data - resultados das múltiplas variáveis categorizadas;
Saída: entropias - valor da entropia.

```
entropia=function(data){
nniveis=sapply(data,nlevels)
osniveis=order(unique(names(rep(nniveis,nniveis))))
dd=rep(osniveis,nniveis)
vv=as.vector(unlist(sapply(data,table)))/nrow(data)
vv[vv==0]=1
entropias=mean(tapply(vv,dd,ent))
list('entropias'=entropias)}
```

A.5 – Obtenção das combinações de n elementos em grupos de k (função interna). É útil para subdividir as categorias das covariáveis categorizadas.

```
combinations=function(n, k){
if(!is.numeric(n) || length(n) != 1 || n%%1) stop("'n' must be an integer")
if(!is.numeric(k) || length(k) != 1 || k%%1) stop("'k' must be an integer")
if(k > n || k <= 0) return(numeric(0))
rowMatrix = function(n) structure(1:n, dim=c(1,n))
colMatrix = function(n) structure(1:n, dim=c(n,1))
if(k == n) return(colMatrix(n))
if(k == 1) return(rowMatrix(n))
L = vector("list", k)
L[[1]] = rowMatrix(2)
L[[2]] = colMatrix(2)
Diff = n-k
for(N in seq(3, n, by=1)){
for(j in seq(min(k, N-1), max(2, N-Diff), by= -1))
    L[[j]] = cbind(L[[j]], rbind(L[[j-1]], N, deparse.level=1))
    if(N <= Diff+1) L[[1]] = rowMatrix(N)
    else L[[N-(Diff+1)]] = numeric(0)
    if(N <= k) L[[N]] = colMatrix(N)
}
L[[k]]
}
```

A.6 – Partição de uma amostra segundo os resultados de uma covariável categorizada (função interna).

Entrada: resp - resultados das variáveis respostas, preditor - resultados da covariável categorizada, nomin - número mínimo de elementos em nós a serem formados, matriz - matriz de dissimilaridades, método - 'entr': entropia e 'sim: dissimilaridade.

Saída: entropias - homog2 - menor medida de heterogeneidade alcançada por uma partição, v - guarda as demais heterogeneidades e particao - partição responsável pela produção de nós menos heterogêneos.

```
partcat=function(resp,preditor,nomin,matriz,método){

v=numeric()
vetora=numeric()
vetorb=numeric()
homog2=numeric()
preditor=factor(preditor)
dados=data.frame(cbind(resp,preditor))
pred2=as.factor(levels(preditor)[table(preditor)!=0])
for (i in 1:trunc(length(pred2)/2)){
      comb=combinations(length(pred2),i)
      for (k in 1:ncol(comb)){
```

```
t=split(dados,dados[,"preditor"]%in% levels(pred2)[comb[,k]])
 if ((nrow(data.frame(t[1]))>=nomin) &&
                (nrow(data.frame(t[2]))>=nomin)){
                  s=t

                  if(método=='sim'){
                        s1=as.data.frame(s[1])
                        s11=as.numeric(row.names(s1))
                        s2=as.data.frame(s[2])
                        s21=as.numeric(row.names(s2))
                        vetora[k]=sum(matriz[s11,s11])
                        vetorb[k]=sum(matriz[s21,s21])
                        homog2[k]=(nrow(s1)/(nrow(s1)+nrow(s2)))*((nrow(s1)*(nrow(s1)-1)/2)^-
1)*vetora[k]+  (nrow(s2)/(nrow(s1)+nrow(s2)))*((nrow(s2)*(nrow(s2)-1)/2)^-1)*vetorb[k]
                  }
                  if(método=='entr'){
                        s1=s[[1]][,1:ncol(resp)]
                        vetora[k]=entropia(s1)$entropias
                        s2=s[[2]][,1:ncol(resp)]
                        vetorb[k]=entropia(s2)$entropias
                        homog2[k]=(nrow(s1)/(nrow(s1)+nrow(s2)))*vetora[k]+
                        (nrow(s2)/(nrow(s1)+nrow(s2)))*vetorb[k]
                  }
            }
            else
            homog2[k]=1000
      }
      v[i]=min(homog2)
      if (v[i]==min(v))
      particao=pred2[comb[,which.min(homog2)]]
}
result=list("homog2"=homog2,"v"=v,"particao"=particao)}
```

A.7 – Partição de uma amostra segundo os resultados de uma covariável numérica (função interna).

Entrada: resp - resultados das variáveis respostas, preditor - resultados da covariável categorizada, nomin - número mínimo de elementos em nós a serem formados, matriz - matriz de dissimilaridades, método - 'entr': entropia e 'sim: dissimilaridade.

Saída: homog2 - a menor medida de heterogeneidade alcançada por uma partição.

```
numer=function(resp,preditor,nomin,matriz,método){
t=1
resp2=(resp[order(preditor),])
preditor=sort(preditor)
resp2=data.frame(resp2)
vetora=rep(1000,nrow(resp)-2*nomin+1)
```

```
vetorb=rep(1000,nrow(resp)-2*nomin+1)
homog2=rep(1000,nrow(resp)-2*nomin+1)
for (i in nomin:(nrow(resp)-nomin)){
      if ((preditor[i]==preditor[i+1])&&(i!=(nrow(resp)-nomin)))
            homog2[t]=1000
      else{
            sa=resp2[1:i,]
            sb=resp2[(i+1):(nrow(resp)),]
            na=nrow(sa)
            nb=nrow(sb)
            razaoa=na/(na+nb)
            razaob=nb/(nb+na)
            if(método=='entr'){
            vetora[t]=entropia(sa)$entropias
            vetorb[t]=entropia(sb)$entropias
            homog2[t]=razaoa*vetora[t]+
            razaob*vetorb[t]}
            if(método=='sim'){
            vetora[t]=sum(matriz[as.numeric(row.names(sa)),as.numeric(row.names(sa))])
            vetorb[t]=sum(matriz[as.numeric(row.names(sb)),as.numeric(row.names(sb))])
            homog2[t]=razaoa*((na*(na-1)/2)^-1)*vetora[t]+
            razaob*((nb*(nb-1)/2)^-1)*vetorb[t]}
      }
      t=t+1
}
p=which(homog2!=1000)[length(which(homog2!=1000))]
q=which(homog2!=1000)[1]
if(sum(preditor[(length(preditor)-nomin+1):length(preditor)]==
preditor[p+nomin-1])!=0)
homog2[p]=1000
if(sum(preditor[1:(nomin-1)]==preditor[q+nomin])!=0)
homog2[q]=1000
result=list("homog2"=homog2)}
```

A.8 – Partição de uma amostra em duas, identificando o tipo de covariável e executando o programa adequado (função interna).

Entrada: resp - resultados das variáveis respostas, pred - resultados de uma covariável, nomin - número mínimo de elementos em nós a serem formados, matriz - matriz de dissimilaridades, método - 'entr': entropia e 'sim: dissimilaridade.

Saída: Medidas de heterogeneidade para o nó pai (homogant), para os nós produzidos (Homog atual), covariáveis, ponto de corte e redução na heterogeneidade (variavel, ponto, reducao), covariáveis, ponto de corte e redução de heterogeneidade da segunda e terceira partições mais importantes (variavel2, ponto2 e reducao2; variavel3, ponto3 e reducao3).

```
partgeral=function(resp,pred,nomin,matriz,método){
op=as.numeric(row.names(resp))
```

```
dados=data.frame(resp,pred)
names(dados)=c(names(resp),names(pred))
vtemp=1000
posicao=rep(0,ncol(pred))
posicao2=rep(0,ncol(pred))
Homog3=numeric()
for (j in 1:ncol(pred)){
      if(is.numeric(pred[,j])){
            if(método=='sim'){
                  a1=numer(resp,pred[,j],nomin,matriz,método='sim')
                  posicao[j]=which.min(a1$homog2)
                  posicao2[j]=which.min(a1$homog2)
            }
            if(método=='entr'){
                  a1=numer(resp,pred[,j],nomin,método='entr')
                  posicao[j]=which.min(a1$homog2)
                  posicao2[j]=which.min(a1$homog2)
            }
      }
      if(is.factor(pred[,j]) && length(nlevels(pred[,j])[table(pred[,j])!=0])>1){
            if(método=='sim')
                  a1=partcat(resp,pred[,j],nomin,matriz,método='sim')
            if(método=='entr')
                  a1=partcat(resp,pred[,j],nomin,método='entr')
            posicao2[j]=as.character(toString(a1$particao))
            if(min(a1$homog2)<vtemp){
                  vtemp=min(a1$homog2)
                  particao=a1$particao
            }
      }

      if(is.factor(pred[,j]) && length(nlevels(pred[,j])[table(pred[,j])!=0])==1)
            Homog3[j]=1000
      else
            Homog3[j]=min(a1$homog2)
}

if(is.numeric(pred[,which.min(Homog3)])){
      corte=c(posicao[which.min(Homog3)],which.min(Homog3))

s=split(cbind(resp,pred),pred[,corte[2]]>pred[order(pred[,corte[2]])[corte[1]+nomin],corte[2]]-
0.000000000001)
      variavel=names(dados)[corte[2]+ncol(resp)]
      ponto=(sort(pred[,corte[2]])[corte[1]+nomin])
}
else if (is.factor(pred[,which.min(Homog3)])){
      s=split(cbind(resp,pred),pred[,which.min(Homog3)]%in% particao)
```

```
        variavel=names(dados)[which.min(Homog3)+ncol(resp)]
        ponto=particao
}
splitalt1=order(rank(Homog3))[2]
splitalt2=order(rank(Homog3))[3]
if(is.numeric(pred[,splitalt1])){
        corte2=c(posicao[splitalt1],splitalt1)
        variavel2=names(dados)[corte2[2]+ncol(resp)]
        ponto2=round((sort(pred[,corte2[2]])[corte2[1]+nomin]),5)
}
else if (is.factor(pred[,splitalt1])){
        variavel2=names(dados)[splitalt1+ncol(resp)]
        ponto2=posicao2[splitalt1]
}
if(is.numeric(pred[,splitalt2])){
        corte3=c(posicao[splitalt2],splitalt2)
        variavel3=names(dados)[corte3[2]+ncol(resp)]
        ponto3=round((sort(pred[,corte3[2]])[corte3[1]+nomin]),5)
}
else if (is.factor(pred[,splitalt2])){
        variavel3=names(dados)[splitalt2+ncol(resp)]
        ponto3=posicao2[splitalt2]
}
if(método=='entr')
        Noint=entropia(resp)$entropias
if(método=='sim')
        Noint=((nrow(resp)*(nrow(resp)-1)/2)^-1)*sum(matriz[op,op])
R=min(Homog3)
R2=sort(Homog3)[2]
R3=sort(Homog3)[3]
reducao=Noint-R
reducao2=Noint-R2
reducao3=Noint-R3
result=list("Homog3"=Homog3,"Homog
atual"=R,"homogant"=Noint,"variavel"=variavel,"ponto"=ponto,"reducao"=round(reducao,5),
'posicao2'=posicao2,'variavel2'=variavel2,'ponto2'=ponto2,'variavel3'=variavel3,'ponto3'=ponto3,'
Reducao2'=reducao2,'Reducao3'=reducao3)}
```

A.9 – Construção e poda da árvore de classificação multivariada.

Entrada: resp - resultados das variáveis respostas, pred - resultados das covariáveis, nomin - número mínimo de elementos em nós a serem formados, ramos - número máximo de níveis da árvore, mincorte - número mínimo de elementos em um nó a ser partido, método - 'entr': entropia e 'sim: dissimilaridade, matriz - matriz de dissimilaridades.

Saída: informações - armazena informações de cada partição, Finais - informações dos nós finais, sumario 2 - tabela de custo complexidade, nosfinais3 - caracterização e outras

características dos nós finais, Compet - segundas e terceiras partições mais importantes em cada passo, dentre outros.

Nota: A árvore de classificação multivariada pode ser encontrada em informações. Em ‘nopai’, têm-se os nós partidos, enquanto os nós originados são armazenados em ‘noesquerdo’ e ‘nodireito’. As partições são executadas das seguintes maneiras: para uma covariável numérica...

```
arvore=function(resp,pred,nomin,ramos,mincorte,método,matriz){
dd=0
for(ç in 1:ncol(pred)){
      if(is.factor(pred[,ç]) | max(table(pred[,ç])>nomin) )
      dd=dd+1
}
cont=1
sussa=list()
sussa2=data.frame()
reducoes2=numeric()
reducoes3=numeric()
red=numeric()
tamanho=numeric()
nosfinais=data.frame(matrix(0,1,2))
nosfinais3=data.frame(matrix(0,1,ncol(resp)+2))
dados=cbind(resp,pred)
r=1
a=numeric()
data=list()
datar=list()
data[[1]]=dados
datar[[1]]=resp
indicavar=numeric()
indicafim=numeric()
pontocat=character()
variavel=numeric()
reducao=vector()
noesquerdo=numeric()
nodireito=numeric()
nopai=numeric()
hant=numeric()
pontonum=numeric()
pontocateg=numeric()

a=seq(1,ramos)
a=2^(a-1)

for (i in 1:(ramos+1)){
      for (k in (2^(i-1)):(2^i-1)){

            dm=as.data.frame(data[k])
```

```
t=rep(0,ncol(pred))
if(dd==ncol(pred)&ncol(dm)>1){
      for(l in 1:ncol(pred))
            t[l]=nrow(dm)-max(table(dm[,ncol(resp)+l]))
}

if(dd!=ncol(pred))
t=nomin+1

t=max(t)
if (nrow(dm)>mincorte && i!=(ramos+1) && t>nomin ){
      names(dm)=names(dados)

      if(método=='sim')

t22=partgeral(dm[,1:ncol(resp)],dm[,(ncol(resp)+1):ncol(dm)],nomin,matriz,método='sim'
)
      if(método=='entr')

t22=partgeral(dm[,1:ncol(resp)],dm[,(ncol(resp)+1):ncol(dm)],nomin,método='entr')
      sussa[[cont]]=c(t22$variavel2,t22$ponto2,t22$variavel3,t22$ponto3)
      sussa2=rbind(sussa2,round(t22$Homog3,5))
      reducoes2=c(reducoes2,t22$Reducao2)
      reducoes3=c(reducoes3,t22$Reducao3)
      cont=cont+1
      variavel[k]=t22$variavel
      pontok=t22$ponto
      reducao[k]=t22$reducao
      hant[k]=t22$homogant
      if (is.numeric(pontok)){
            dados2=split(dm,dm[,variavel[k]]>(pontok-0.000001))[1]
            dados3=split(dm,dm[,variavel[k]]>(pontok-0.000001))[2]
            pontocat2=as.character(round(pontok,5))
            pontonum=c(pontonum,pontok)
            pontocateg=c(pontocateg,NA)
            indicavar=c(indicavar,"n")
      }
      if (is.factor(pontok)){
            dados2=split(dm,dm[,variavel[k]]%in%pontok)[1]
            dados3=split(dm,dm[,variavel[k]]%in%pontok)[2]
            pontocat2=toString(as.character(pontok))
            pontocateg=c(pontocateg,toString(as.character(pontok)))
            pontonum=c(pontonum,NA)
            indicavar=c(indicavar,"c")
      }
      dados2=as.data.frame(dados2)
      dados3=as.data.frame(dados3)
```

```
        names(dados2)=names(dados)
        names(dados3)=names(dados)
        data[[r+1]]=dados2
        data[[r+2]]=dados3
        datar[[r+1]]=dados2[,1:ncol(resp)]
        datar[[r+2]]=dados3[,1:ncol(resp)]
        r=r+2
        pontocat=c(pontocat,pontocat2)
}

else{
        if(nrow(dm)>=nomin)
        indicafim=c(indicafim,k)
        data[[r+1]]=1
        data[[r+2]]=2
        datar[[r+1]]=1
        datar[[r+2]]=2
        r=r+2
        variavel[k]=NA
        reducao[k]=NA
        pontocat[k]=NA
        hant[k]=0
}

t=as.numeric(row.names(dm))
if(método=='sim')
        hh=sum(matriz[t,t])/(length(t)*(length(t)-1)/2)

if(método=='entr'){
        if(ncol(data.frame(dm))>1)
                hh=entropia(dm[,1:ncol(resp)])$entropias
        else
                hh=1000
}

tt=numeric()
for (e in 1:ncol(dm))
        tt[e]=(names(which.max(table(dm[,e]))))

nosfinais=rbind(nosfinais,c(k,hh))
tamanho=c(tamanho,nrow(dm))

nosfinais3=rbind(nosfinais3,c(k,hh,tt))

nopai[k]=k
noesquerdo[k]=r-1
nodireito[k]=r
```

```
        }
}
nosfinais=nosfinais[-1,]
nosfinais3=nosfinais3[-1,]
nosfinais2=nosfinais[indicafim,]
nosfinais4=nosfinais3[indicafim,]
tamanho2=tamanho[indicafim]
onais=data.frame(na.omit(cbind(nosfinais2,tamanho2)))
nosdecima=trunc(onais[,1]/2)
hant2=hant[nosdecima]
onais=na.omit(cbind(onais,nosdecima,hant2))
informações=na.omit(data.frame(nopai,noesquerdo,nodireito,variavel,pontocat,reducao))
onais=cbind(onais,rep(0,nrow(onais)))
for (j in 1:nrow(onais))
onais[j,6]=informações[which(informações[,1]==onais[j,4]),6]
names(onais)=c("Número","Homogeneidade","Tamanho","Nó Pai","Homog Ant","Redução")
d=list()
w=2
Finais=onais
homogf=sum(Finais[,2]*Finais[,3])/nrow(dados)
d[[1]]=Finais
red[1]=sum(Finais[,2]*Finais[,3])/nrow(resp)
nnos=nrow(Finais)
seque=seq(nnos,1)
while(nrow(Finais)>1){
        Finais2=Finais[which(Finais[,4]%in%as.numeric(names(which(table(Finais[,4])==2)))),]
        Finais2=Finais2[order(Finais2[,6]),]
        Finais1=Finais[which(Finais[,4]%in%as.numeric(names(which(table(Finais[,4])==1)))),]
        Finais=rbind(Finais2,Finais1)
        posição=which(informações[,2]==Finais[1,4] | informações[,3]==Finais[1,4])

Finais=rbind(Finais,c(Finais[1,4],Finais[1,5],Finais[1,3]+Finais[2,3],informações[posição,1],nosf
inais[,2][which(nosfinais[,1]==informações[posição,1])],informações[posição,6]))
        Finais=Finais[-c(1,2),]
        d[[w]]=Finais
        d2=data.frame(d[w])
        red[w]=sum(d2[,2]*d2[,3])/nrow(resp)
        Finais=Finais
        w=w+1
}
custocomp=data.frame(cbind(seq(1,nrow(data.frame(d[1]))),red[order(seque)]))
mat=custocomp
sumario=data.frame(matrix(0,1,2))
sumario2=data.frame(matrix(0,1,2))
alfa=0.0001
g=2
while(sumario[g-1,2]!=1){
```

```
        ralfa=mat[,2]+alfa*mat[,1]
        sumario=rbind(sumario,c(alfa,which.min(ralfa)))
        if(sumario[g,2]!=sumario[g-1,2])
              sumario2=rbind(sumario2,c(alfa,which.min(ralfa)))
        g=g+1
        alfa=alfa+0.0001
}
sumario=sumario[-1,]
sumario2=sumario2[-1,]
s3=numeric()
for (t in 1:nrow(sumario2))
        s3=c(s3,mat[,2][which(mat[,1]==sumario2[t,2])])
errorel=s3/s3[length(s3)]
sumario2=cbind(sumario2,s3,errorel)
names(sumario2)=c("Alfa","Tamanho","R(alfa)","Erro relativo")
p=length(sumario2[,1])
if(método=='entr'){
        taman=lapply(datar,length)
        indicanos=nosfinais3[which(taman>1),1]
        indicanos=as.numeric(indicanos)
        nosfinais3=nosfinais3[indicanos,]
        nosfinais3=nosfinais3[,-2]
}
if(método=='sim'){
        nosfinais3=nosfinais3[which(nosfinais3[,2]!=NaN),]
        nosfinais3=nosfinais3[,-2]
        indicanos=as.numeric(nosfinais3[,1])
}
alfag=sqrt(sumario2[,1][1:(p-1)]*sumario2[,1][2:p])
alfag=c(alfag,1)
s= lapply(datar[indicanos],entropia)
Entropias=as.vector(unlist(s))
Tamanho=tamanho[which(tamanho!=1)]
nosfinais3=cbind(nosfinais3,Entropias,Tamanho)
nosfinais4=nosfinais3[which(nosfinais3[,1]%in%indicafim),]
sussa=matrix(unlist(sussa),length(indicavar),4,byrow=T)
S2=apply(sussa2,1,sort)[2,]
S3=apply(sussa2,1,sort)[3,]
sussa=data.frame(sussa[,1],sussa[,2],round(reducoes2,5),sussa[,3],sussa[,4],round(reducoes3,5))
names(sussa)=c('Var-2ºCorte','Pt-2ºCorte','Hg-2ºCorte','Var-3ºCorte','Pt-3ºCorte','Hg-3ºCorte')
names(sussa2)=names(pred)
result=list("informações"=cbind(informações,indicavar,pontocateg,pontonum)
,"Finais"=onais,"sumario2"=sumario2,"nosfinais3"=nosfinais3,'nomin'=nomin,'ramos'=ramos,'mi
ncorte'=mincorte,'Compet'=sussa,"data"=data,"d"=d,"alfag"=alfag)}
```

A.10 – Predição/classificação de novos elementos usando uma árvore de classificação multivariada.

Entrada: datapred - vetores de covariáveis dos elementos, inf (opcional) - informações do modelo construído e modelo - um objeto tipo arvore.

Saída: numero – números dos nós aos quais os elementos são alocados, class – classificações dos elementos, datapred – vetores de covariáveis.

```
predic=function(datapred,inf,modelo){
if(missing(inf))
inf=modelo$informações
inf1=cbind(inf[which(inf[,7]=="n"),],(modelo$informações[,'pontonum']
[which(inf[,7]=="n")]))
inf2=cbind(inf[which(inf[,7]=="c"),],(modelo$informações[,'pontocateg']
[which(inf[,7]=="c")]))

names(inf1)=c(names(inf),"pnum")
names(inf2)=c(names(inf),"pcat")
t2=numeric()
p2=numeric()
s=matrix(1,nrow(datapred),modelo$ramos+2)

for (k in 1:(modelo$ramos+1)){
      for (l in 1:nrow(datapred)){
            noat=s[l,k]
            if(noat%in%inf[,1]){
                  t=(inf[,4][which(inf[,1]==noat)])
                  w=which(names(datapred)==t)
                  if (is.numeric(datapred[,w])){
                        p=inf1[,9][which(inf1[,1]==noat)]
                        if(datapred[l,w]<=(p-0.0000001))
                              s[l,k+1]=2*s[l,k]
                        if(datapred[l,w]>(p-0.000001))
                              s[l,k+1]=2*s[l,k]+1
                  }
                  else {
                        p=inf2[,8][which(inf2[,1]==s[l,k])]
                        if(datapred[l,w]%in% strsplit(as.character(p),split=", ")[[1]])
                        s[l,k+1]=2*s[l,k]+1
                        else
                        s[l,k+1]=2*s[l,k]
                  }
            }
            else
            s[l,k+1]=1
            t2=c(t2,t)
            p2=c(p2,p)
      }
```

```
}
modelo$nosfinais3[,1]=as.numeric(modelo$nosfinais3[,1])
class=data.frame()
numero=numeric()
for(i in 1:nrow(s)){
      numero[i]=max(s[i,])

class=rbind(class,modelo$nosfinais3[which(modelo$nosfinais3[,1]==numero[i]),2:(ncol(resp)+2-
1)])
}
list("numero"=numero,"class"=class,"datap"=datapred)}
```

A.11 - Predição/classificação de novos elementos, utilizada no processo de validação cruzada (função interna).

Entrada: datapred e dataresp- vetores de covariáveis e vetores de resposta dos elementos a serem classificados, inf e nf -informações das partições e dos nós finais da árvore e modelo - modelo utilizado para classificação.

Saída: taxac - taxa de classificações coorretas por validação cruzada, class e número – classificações e números dos nós aos quais os elementos são alocados.

```
predt2=function(datapred,dataresp,inf,nf,modelo){

inf1=cbind(inf[which(inf[,7]=="n"),],(modelo$informações[,'pontonum']
[which(inf[,7]=="n")]))
inf2=cbind(inf[which(inf[,7]=="c"),],(modelo$informações[,'pontocateg']
[which(inf[,7]=="c")]))
names(inf1)=c(names(inf),"pnum")
names(inf2)=c(names(inf),"pcat")

t2=numeric()
p2=numeric()

datapred=data.frame(datapred)
dataresp=data.frame(dataresp)

s=matrix(1,nrow(datapred),modelo$ramos+2)

for (k in 1:(modelo$ramos+1)){
      for (l in 1:nrow(datapred)){
            noat=s[l,k]
            if(noat%in%inf[,1]){
                  t=(inf[,4][which(inf[,1]==noat)])
                  w=which(names(datapred)==t)
                  if (is.numeric(datapred[,w])){
                        p=inf1[,9][which(inf1[,1]==noat)]
                        if(datapred[l,w]<=(p-0.0000001))
```

```
                        s[l,k+1]=2*s[l,k]
                    if(datapred[l,w]>(p-0.0000001))
                        s[l,k+1]=2*s[l,k]+1
                }
                else {
                    p=inf2[,8][which(inf2[,1]==s[l,k])]
                    if(datapred[l,w]%in% strsplit(as.character(p),split=", ")[[1]])
                        s[l,k+1]=2*s[l,k]+1
                    else
                        s[l,k+1]=2*s[l,k]
                }
            }
            else
                s[l,k+1]=1
            t2=c(t2,t)
            p2=c(p2,p)
        }
}
nf[,1]=as.numeric(nf[,1])
numero=numeric()
class=data.frame()
numero=numeric()
for(i in 1:nrow(s)){
    numero[i]=s[i,max(which(s[i,]%in%nf[,1]))]
    class=rbind(class,nf[which(nf[,1]==numero[i]),2:(ncol(dataresp)+2-1)])
    numero=c(numero,nf[which(nf[,1]==numero[i]),1])
}
ac=sum(class==dataresp)
er=sum(class!=dataresp)
taxac=ac/(ac+er)
list("taxac"=taxac,"class"=class,"numero"=numero)}
```

A.12 - Validação cruzada.

Entrada: pred - vetores de covariáveis, resp - vetores de respostas, modelo - objeto do tipo arvore, alfag - (opcional) - valores para o parâmetro de complexidade, xval - número de validações cruzadas e método - 'entr' ou 'sim'.

Saída: acertos, Distância e Entropia - calculados para cada valor do parâmetro de complexidade em cada passo da validação cruzada, real - dados observados, preditos - classificações.

```
vcross=function(pred,resp,modelo,alfag,xval,método){
if(missing(alfag))
alfag=modelo$alfag
y=0
acertos=matrix(0,length(alfag),xval)
ddistancia=matrix(0,length(alfag),xval)
```

```
matentrop=matrix(0,length(alfag),xval)
xg=sample(rep(1:xval,length=nrow(resp)),nrow(resp),replace=F)
real=list()
predito=list()
for(d in 1:length(alfag)){
      real[[d]]=data.frame(resp[1,])
      predito[[d]]=data.frame(resp[1,])
      names(real[[d]])=names(resp)
      names(predito[[d]])=names(resp)}
      for(i in 1:xval){
            nn=which(xg==i)
            datapred=pred[which(xg==i),]
            dataresp=resp[which(xg==i),]
            datapredconst=pred[-which(xg==i),]
            datarespconst=resp[-which(xg==i),]
            if(método=='sim')

modeloc=arvore(datarespconst,datapredconst,nomin=modelo$nomin,ramos=modelo$ramos,minc
orte=modelo$mincorte,matriz=matriz,método='sim')
            if(método=='entr')

modeloc=arvore(datarespconst,datapredconst,nomin=modelo$nomin,ramos=modelo$ramos,minc
orte=modelo$mincorte,método='entr')
            precisa=list()
            for (k in 1:length(alfag)){
                  if(alfag[k]<modeloc$sumario[nrow(modeloc$sumario[1]),1]){
                        a=which(modeloc$sumario[1]>alfag[k])[1]
                  if(a!=1)
                        a=a-1
                  tamanho=modeloc$sumario[a,2]
                  b=numeric()
                  for(t in 1:length(modeloc$d)){
                  if(nrow(data.frame(modeloc$d[t]))==tamanho)
                        b[t]=1
                  else
                        b[t]=0}

                  a=modeloc$d[[which(b==1)]][,1]

                  a1=matrix(0,length(a),modelo$ramos+2)
                  a1[,1]=a
                  for (q in 1:(modelo$ramos+1))
                        a1[,q+1]=trunc(a1[,q]/2)
                  a1=unique(as.vector(a1))
                  a1=a1[which(a1!=0)]
                  precisa[[k]]=modeloc$informações[which(modeloc$informações
                  [,2]%in%a1 | modeloc$informações
```

```
                [,3]%in%a1),]

                p=data.frame(precisa[k])

                if(sum(p[-which(p[,2]%in%p[,1]),2])==0)
                      indicanosfinaispar=p[,2]
                if(sum(p[-which(p[,2]%in%p[,1]),2])!=0)
                      indicanosfinaispar=p[,2][-which(p[,2]%in%p[,1])]

                if(sum(p[-which(p[,3]%in%p[,1]),3])==0)
                      indicanosfinaisimpar=p[,3]
                if(sum(p[-which(p[,3]%in%p[,1]),3])!=0)
                      indicanosfinaisimpar=p[,3][-which(p[,3]%in%p[,1])]

                indicanosfinais=c(indicanosfinaispar,indicanosfinaisimpar)


nosfinais3ag=modeloc$nosfinais3[which(modeloc$nosfinais3[,1]%in%indicanosfinais),]

                caminho=matrix(1,length(indicanosfinais),8)
                caminho[,1]=indicanosfinais
                e=1
                for (w in 1:length(indicanosfinais)){
                      while(caminho[w,e]!=1){
                            caminho[w,e+1]=trunc(caminho[w,e]/2)
                            e=e+1
                      }
                      e=1
                }

                cam2=unique(as.vector(caminho))
                inf2=modeloc$informações[which(modeloc$informações[,2]%in%cam2 |
modeloc$informações[,3]%in%cam2),]

                if (nrow(inf2)>1){
                      indicanosfinais2=c(inf2[-which(inf2[,2]%in%inf2[,1]),2],inf2[-
which(inf2[,3]%in%inf2[,1]),3])

idc=modeloc$nosfinais3[which(modeloc$nosfinais3[,1]%in%indicanosfinais),]
                      matentrop[k,i]=sum(idc[,'Entropias']*idc[,'Tamanho'])/nrow(resp)
                }

                if (nrow(inf2)==1)

matentrop[k,i]=sum(modeloc$nosfinais3[2:3,'Entropias']*modeloc$nosfinais3[2:3,'Tamanho'])/nr
ow(resp)
```

```
                predt=predt2(datapred,dataresp,inf=inf2,nf=nosfinais3ag,modelo=modeloc)
                matpred=predt$class
                names(matpred)=names(resp)
                acertos[k,i]=predt$taxac
                numero=predt$numero
                }
                else{
                        matpred=data.frame(resp[1,])
                        for (u in 1:ncol(dataresp))

matpred[1,u]=levels(datarespconst[,u])[which.max(table(datarespconst[,u]))]
                        for(o in 1:(nrow(dataresp)-1))
                                matpred=rbind(matpred,matpred[1,])
                        acertos[k,i]=sum(matpred==dataresp)/(nrow(dataresp)*ncol(dataresp))

                        numero=rep(1,length(nn))

matentrop[k,i]=sum(modeloc$nosfinais3[1,ncol(resp)+2]*modeloc$nosfinais3[1,ncol(resp)+3])/n
row(resp)
                }
                distancia=numeric()
                for(h in 1:nrow(datapred)){
                        data1=data.frame(modeloc$data[numero[h]])
                        data2=as.numeric(rownames(data1))

                        if(alfag[k]<modeloc$sumario[nrow(modeloc$sumario[1]),1])
                                distancia[h]=sum(matriz[nn[h],data2])/length(data2)

                        else
                                distancia[h]=sum(matriz[nn[h],1:nrow(resp)])/nrow(resp)
                }
                ddistancia[k,i]=mean(distancia)
                ag=alfag[k]
                y=y+1
                predito[[k]]=rbind(predito[[k]],matpred)
                real[[k]]=rbind(real[[k]],dataresp)
        }
}
for (i in 1:length(alfag)){
        predito[[i]]=predito[[i]][-1,]
        real[[i]]=real[[i]][-1,]
}
medacertos=rowSums(acertos)/xval
list("acertos"=acertos,"Distâncias"=ddistancia,"Entropias"=matentrop,"real"=real,"predito"=predi
to)}
```

A.13 – Seleção do modelo, com base em novas validações cruzadas e na regra do desvio padrão.

Entrada: modelo - um objeto do tipo arvore, vcruz - um objeto do tipo vcross e método - 'entr' ou 'sim'.

Saída: Predições – classificações dos elementos para o modelo selecionado, Respostas – vetores observados de respostas, Árvore – informações das partições do modelo, k – número de nós finais.

```
seleção=function(modelo,vcruz,método){
m=rowSums(vcruz$Distância)/ncol(vcruz$Distância)
a=data.frame(modelo$sumario[,c(1,2,4)],rowMeans(vcruz$Distância)/max(rowMeans(vcruz$Dis
tância)))
names(a)=c("CP", "nsplit", "rel error", "xerror")
tam=modelo$sumario2[,2]
alfagg=a[which.min(a[,4]),1]
f=numeric()
e=numeric()
for (t in 1:10){
      set.seed(runif(1,1,2000))
      r=vcross(pred, resp, modelo,alfagg, xval=10,método='entr')
      f=rbind(f,(r$Distância))
      e=rbind(e,(r$Entropias))
}

gráfico=plot(tam,m,type="b",xlab="Tamanho",ylab="Dissimilaridade média")
e2=rowMeans(e)
f2=rowMeans(f)
lines(c(0,max(tam)),c(min(m)+sd(f2),min(m)+sd(f2)),lty=2,col="red")
title("Gráfico de custo complexidade")
points(a[max(which(m<min(m)+sd(f2))),2],m[max(which(m<min(m)+sd(f2)))],col="red",pch=2
0,cex=2.0)
points(a[which.min(m),2],m[which.min(m)],col="blue",pch=20,cex=2.0)
gráfico
k=min(a[,2][which(m<min(m)+sd(f2))])
indicalista=which(a[,2]==k)
predições=vcruz$predito[[indicalista]]
predições=predições[-1,]
respostas=vcruz$real[[indicalista]]
respostas=respostas[-1,]
cap=list()
tabelas=list()
for(u in 1:ncol(respostas)){
      tabelas[[u]]=table(predições[,u],respostas[,u])
      cap[[u]]=diag(tabelas[[u]]/rowSums(tabelas[[u]]))
}
a=data.frame(modelo$d[length(modelo$d)-k+1])
indicanosfinais=a[,1]
caminho=matrix(1,length(indicanosfinais),8)
```

```
caminho[,1]=indicanosfinais
e2=1
for (w in 1:length(indicanosfinais)){
      while(caminho[w,e2]!=1){
            caminho[w,e2+1]=trunc(caminho[w,e2]/2)
            e2=e2+1
      }
      e2=1
}
cam2=unique(as.vector(caminho))
inf2=modelo$informações[which(modelo$informações[,2]%in%cam2 |
modelo$informações[,3]%in%cam2),]
list("Predições"=predições,"Respostas"=respostas,"Árvore"=inf2,"k"=k)}
```

A.14 – Seleção de uma árvore de tamanho k, escolhido pelo usuário.

Entrada: modelo - objeto do tipo arvore, vcruz - objeto do tipo vcross, método - 'entr' ou 'sim' e k - tamanho da árvore.

Saída: Predições e respostas dos elementos, Árvore - informações das partições e k - tamanho da árvore.

```
seleção2=function(modelo,vcruz,método,k){
if(método=='sim'){
      m=rowSums(vcruz$Distância)/ncol(vcruz$Distância)

a=data.frame(modelo$sumario[,c(1,2,4)],rowMeans(vcruz$Distância)/max(rowMeans(vcruz$Dis
tância)))
      names(a)=c("CP", "nsplit", "rel error", "xerror")
}
if(método=='entr'){
      m=rowSums(vcruz$acertos)/ncol(vcruz$acertos)

a=data.frame(modelo$sumario[,c(1,2,4)],rowMeans(vcruz$acertos)/max(rowMeans(vcruz$acerto
s)))
      names(a)=c("CP", "nsplit", "rel error", "xerror")
}
tam=modelo$sumario2[,2]
indicalista=which(a[,2]==k)
predições=vcruz$predito[[indicalista]]
respostas=vcruz$real[[indicalista]]
cap=list()
tabelas=list()
for(u in 1:ncol(respostas)){
      tabelas[[u]]=table(predições[,u],respostas[,u])
      cap[[u]]=diag(tabelas[[u]]/rowSums(tabelas[[u]]))
}
a=data.frame(modelo$d[length(modelo$d)-k+1])
```

```
indicanosfinais=a[,1]
caminho=matrix(1,length(indicanosfinais),8)
caminho[,1]=indicanosfinais
e2=1
for (w in 1:length(indicanosfinais)){
      while(caminho[w,e2]!=1){
            caminho[w,e2+1]=trunc(caminho[w,e2]/2)
            e2=e2+1
      }
      e2=1
}
cam2=unique(as.vector(caminho))
inf2=modelo$informações[which(modelo$informações[,2]%in%cam2 |
modelo$informações[,3]%in%cam2),]
list("Predições"=predições,"Respostas"=respostas,"Árvore"=inf2,"k"=k)
}
```

A.15 – Extração de resultados do modelo selecionado.

Entrada: modelo - objeto do tipo arvore, vcruz - objeto do tipo vcross, método - 'entr' ou 'sim', sel - objeto do tipo seleção (caso se queira o resumo da árvore selecionada pela regra do desvio padrão) e tam - tamanho da árvore (caso se queira o resumo de uma árvore de tamanho especificado.

Saída: arvore final - partições da árvore selecionada, Entropia, Erros e Dissimilaridades - medidas para o modelo selecionado, dataprop - frequências dos nós finais quanto às variáveis resposta, tabs - tabelas de classificado versus observado, freq - frequencias de elementos para cada variável resposta,

```
resumao=function(modelo,vcruz,método,sel,tam){

Erros=(1-rowMeans(vcruz$acertos))
Erros=(Erros-mean(Erros))/(sd(Erros))
Erros2=rowMeans(1-vcruz$acertos)
Entropias=rowMeans(vcruz$Entropias)
Entropias=(Entropias-mean(Entropias))/(sd(Entropias))
Entropias2=rowMeans(vcruz$Entropias)
Distâncias=rowMeans(vcruz$Distâncias)
Distâncias=(Distâncias-mean(Distâncias))/(sd(Distâncias))
Distâncias2=rowMeans(vcruz$Distâncias)
Tamanho=modelo$sumario2[,2]
Tamanho2=Tamanho
Tamanho=(Tamanho2-mean(Tamanho2))/(sd(Tamanho2))

par(mfrow=c(1,3))

plot(Tamanho2,Erros2,type='b',xlab='Tamanho',ylab='Taxa de erros')
```

```
lines(c(Tamanho2[1],Tamanho2[length(Tamanho2)]),c(Erros2[1],Erros2[length(Erros2)]),col='red',lty=2)
plot(Tamanho2,Distâncias2,type='b',xlab='Tamanho',ylab='Dissimilaridade média')
lines(c(Tamanho2[1],Tamanho2[length(Tamanho2)]),c(Distâncias2[1],Distâncias2[length(Distâncias2)]),col='red',lty=2)
plot(Tamanho2,Entropias2,type='b',xlab='Tamanho',ylab='Entropia')
lines(c(Tamanho2[1],Tamanho2[length(Tamanho2)]),c(Entropias2[1],Entropias2[length(Entropias2)]),col='red',lty=2)
mtext("Ajuste", outer = TRUE, line = -2.5,cex=1.5)

if(missing(tam)){

z=which(Tamanho2==sel$k)
distanc=Distâncias2[z]
acert=Erros2[z]
k=Tamanho2[z]
}
else{
z=which(Tamanho2==tam)
distanc=Distâncias2[z]
acert=Erros2[z]
k=Tamanho2[z]
}
suma=modelo$sumario
avre=seleção2(modelo,vcruz,método,k)
t=1
w=c(avre$Árvore[,2],avre$Árvore[,3])
if(k==1)
      w=1
if(length(w)>2)
      w=w[-which(w%in%avre$Árvore[,1])]
w=sort(w)
nresp=ncol(vcruz$predito[[1]])
nlev=sapply(vcruz$predito[[1]],nlevels)
dataprop=numeric()
for(q in 1:length(w)){
      vec=numeric()
      for(z in 1:nresp)
      vec=c(vec,table(modelo$data[[w[q]]][,z]))
      dataprop=rbind(dataprop,vec)
}
d=list()
e=list()
for(i in 1:length(w)){
      for(l in 1:nresp){
```

```
d[[t]]=c(w[i],nrow(modelo$data[[w[i]]]),names(which.max(table(modelo$data[[w[i]]][,l]))),roun
d(max(table(modelo$data[[w[i]]][,l]))/
            (sum(table(modelo$data[[w[i]]][,l]))),2))
            t=t+1
      }
}
s=data.frame(matrix(unlist(d),k*ncol(vcruz$real[[1]]),4,byrow=T))
names(s)=c('Nó',"Tamanho",'Clasel2e','Proporção')
nomesc=as.vector(sapply(vcruz$predito[[1]][,1:nresp],levels))
vars=paste('v',seq(1,nresp),sep='')
vars2=rep(vars,as.numeric(nlev))
vars3=paste(vars2,unlist(nomesc),sep=':')
nomes=paste('N',w,sep="")
rownames(dataprop) <- nomes
colnames(dataprop)=vars3
tabs=list()
preditos=list()
real=list()
var=list()
diags=list()
for(i in 1:nresp){
      preditos[[i]]=avre$Predições[,i]
      real[[i]]=avre$Respostas[,i]
      Real=real[[i]]
      Predito=preditos[[i]]
      tabs[[i]]=table(Real,Predito)
      tabs[[i]]=round(tabs[[i]]/rowSums(table(real[[i]],preditos[[i]])),2)
      var[[i]]=table(Real)
      diags[[i]]=diag(tabs[[i]])
}
diags=round(unlist(var)*unlist(diags),0)
datacert=rbind(diags,unlist(var)-diags)
par(mfrow=c(1,nresp))
k2=0
if(k!=1){
      for(i in 1:nresp){

barplot(datacert[1:2,(k2+1):(k2+nlev[[i]])],col=c('blue','red'),xlab=paste('Variável',i),legend=c('A
certos','Erros'))
            k2=k2+nlev[[i]]
      }
}
colnames(datacert)=vars3
k2=0
par(mfrow=c(1,nresp))
if(k!=1){
```

```
	for(i in 1:nresp){
		barplot(t(dataprop[1:length(w),(k2+1):(k2+nlev[[i]])]),col=hcl(h = seq(0,270, length =
nlev[[i]])),xlab=paste('Variável',i),legend=colnames(dataprop)[(k2+1):(k2+nlev[[i]])])
		k2=k2+nlev[[i]]}
	}
en=numeric()
tam=numeric()
if(k!=1){
	for (i in 1:length(w)){
		nof=modelo$data[[w[i]]][,1:nresp]
		en[i]=entropia(nof)$entropias
		tam[i]=nrow(nof)
	}
}
if(k==1){
	tam=nrow(modelo$data[[1]])
	en=entropia(modelo$data[[1]][,1:nresp])$entropias
}
entropianova=sum((en*tam)/sum(tam))
vetent=numeric()
nnos=as.numeric(lapply(modelo$d,nrow))
for(s in 1:length(Tamanho2)){
	nfim=modelo$d [[which(nnos==Tamanho2[s])]][,1]
	tfim=as.numeric(lapply(modelo$data[nfim],nrow))
	entfim=numeric()
	for(y in 1:length(nfim))
	entfim[y]=entropia(modelo$data[[nfim[y]]][,1:ncol(vcruz$real[[1]])])$entropia
	vetent[s]=sum((entfim*tfim)/sum(tfim))

}
rownames(datacert)=c('Acertos','Erros')
list('arvore final'=avre[[3]],'Entropia'=entropianova,'Erros'=1-acert,'Dissimilaridades'=distanc
,'dataprop'=dataprop,'tabs'=tabs,'var'=var,'datacert'=datacert)
}
```