BRUNO CAETANO VIDIGAL

AVALIAÇÃO DE AGRUPAMENTOS EM MISTURA DE VARIÁVEIS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2013

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV

T
V653a
2013

Vidigal, Bruno Caetano, 1988-
Avaliação de agrupamentos em mistura de variáveis / Bruno Caetano Vidigal. – Viçosa, MG, 2013.
xi, 56f. : il. ; 29cm.

Inclui apêndices.
Orientador: Paulo Roberto Cecon
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 40-44

1. Análise multivariada. 2. Análise por agrupamento. 3. Genômica. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.535

BRUNO CAETANO VIDIGAL

AVALIAÇÃO DE AGRUPAMENTOS EM MISTURA DE VARIÁVEIS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

Aprovada em 6 de fevereiro de 2013.

Adésio Ferreira	Moysés Nascimento (Coorientador)

Paulo Roberto Cecon
(Orientador)

À minha mãe

"Tudo o que um sonho precisa para ser realizado é alguém que acredite que ele possa ser realizado"

Roberto Shinyashiki

"To infinity ... and beyond"

Buzz Lightyear

AGRADECIMENTOS

Agradeço primeiramente a Deus por sempre me iluminar, me dando muita sabedoria, tranqüilidade e paz.

À minha mãe que sempre fez o impossível por mim. Muito obrigado por todos ensinamentos, carinho e amor. Agradeço também a toda minha família.

À Universidade Federal de Viçosa e à CAPES pelo financiamento de minha bolsa de estudo.

Ao professor e orientador Paulo Roberto Cecon que me incentivou. Ao professor e co-orientador Moysés Nascimento que me deu muita força em minha pesquisa, sempre com paciência e boa vontade em ajudar. Ao professor e co-orientador Cosme Damião Cruz, que também me ajudou bastante a conseguir concluir esse trabalho e ao professor Adésio Ferreira que contribuiu com ótimas sugestões.

Aos amigos do mestrado Diego, Cássio, Wagner, Vinícius, Fátima, Pâmela, Camila, Renata, Márcio, Édimo, Nayara, Leillimar e também aos integrantes da República Os Pirigozo Silvano, Fernandin, Franklin e Evaldo. Todos vocês fizeram parte da minha família aqui em Viçosa. Vou sentir falta demais das resenhas no laboratório, dos churrascos e festas. Foi tudo muito bom.

Em especial eu tenho que falar do Wagner, que é um grande amigo e também do Cássio, que é um dos caras mais chatos que conheço.

Aos outros amigos que fiz por aqui, como a Karla, Fabiene, às meninas da República D.I.V.A.S, Martijn, Leon, Nívea.

Aos grandes amigos do tempo de graduação que volta e meia freqüentavam minha casa em Viçosa, sempre trazendo muita alegria e ótimas histórias. Valeu Samuel, Luís, Iago, Priscila, Roberto e Laura.

Enfim, agradeço a todos que me ajudaram não só a fazer esse trabalho, mas que me acompanharam nessa fase que estou concluindo em minha vida.

Um beijo e um sorriso a todos.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

VIDIGAL, Bruno Caetano, M.Sc., Universidade Federal de Viçosa, fevereiro de 2013. **Avaliação de Agrupamentos em mistura de variáveis**. Orientador: Paulo Roberto Cecon. Coorientadores: Moysés Nascimento e Cosme Damião Cruz.

A análise de agrupamento é amplamente utilizada em muitas áreas de pesquisa a fim de se reconhecer uma estrutura padrão de variabilidade entre os indivíduos ou objetos estudados, classificando-os em grupos homogêneos. No entanto, dos trabalhos publicados, a maioria deles versam apenas sobre variáveis numéricas, excluindo da análise, as informações contidas nas variáveis categóricas. Dessa forma, esse trabalho teve o objetivo de avaliar várias formas de agrupamentos em um banco de dados simulado e também de disponibilizar uma rotina em R do algoritmo k-protótipos e uma rotina para se realizar agrupamentos hierárquicos. As medidas de distâncias avaliadas foram: euclidiana, euclidiana ao quadrado, euclidiana média, mahalanobis, manhattan, medidas combinadas e a de gower. Quanto aos algoritmos de agrupamento hierárquicos utilizados foram: vizinho mais próximo, vizinho mais distante, UPGMA e ward . Os algoritmos não-hierárquicos foram: k-médias e o k-protótipos. Os resultados obtidos foram confrontados entre si e concluiu-se que os algoritmos não-hierárquicos foram superiores aos hierárquicos e que incluir variáveis categóricas na análise é viável.

ABSTRACT

VIDIGAL, Bruno Caetano, M.Sc., Universidade Federal de Viçosa, February, 2013. **Evaluation of Cluster Variables in a Mixture**. Advisor: Paulo Roberto Cecon. Co-advisors: Moysés Nascimento and Cosme Damião Cruz.

Cluster analysis is widely used in many research areas in order to recognize a standard structure of variability between individuals or objects studied, classifying them into homogeneous groups. However, the studies that are published, most of them deal only on numeric variables, excluding the analysis, the information contained in categorical variables. Thus, this study aims to evaluate some similarity measures and clustering algorithms in databases and also simulated on a case study in Genetics. The similarity measures evaluated were: euclidean, squared euclidean, mean euclidean, mahalanobis, manhattan, combined measures and gower. The hierarchical clustering algorithms are: nearest neighbor, furthest neighbor, UPGMA and Ward. The algorithms evaluated from the class of non-hierarchical are the k-means and k-prototypes, which is an extension of the first. The results were compared and we concluded the non-hierarquical were better than hierarquical methods.

1. INTRODUÇÃO

Conhecer o comportamento de indivíduos e/ou objetos e a forma como eles estão dispostos em meio às tantas variáveis que os caracterizam faz com que seja de suma importância explorar métodos estatísticos multivariados.

Dentro da estatística multivariada, a análise de agrupamento (*cluster*) ocupa um papel fundamental que é o de alocar indivíduos e/ou objetos em grupos de acordo com as variáveis que foram medidas e avaliadas.

A análise de agrupamento está presente em estudos que envolvem educação (PIMENTEL et al, 2003), economia (SIMÕES, 2003), análise de crédito (BRITO et al, 2009), experimentos agronômicos (TOTTI et al 2001), melhoramento genético (ELIAS et al, 2007), dentre outros.

Mais especificamente, dentro do melhoramento genético, existe a diversidade genética que, por exemplo, têm o objetivo de direcionar cruzamentos de indivíduos e isso pode ser feito através da análise de agrupamento (CRUZ et al., 2008).

O estudo da diversidade genética é de vital importância para o melhoramento genético, pois dá condições ao pesquisador de definir uma estratégia de seleção em busca dos melhores genes (FERREIRA, 2007).

Muitos trabalhos acadêmicos envolvem somente a utilização de variáveis numéricas quando se trata de agrupamento. Porém existem métodos que conseguem trabalhar com mistura de variáveis, ou seja, envolvendo tanto variáveis numéricas quanto categóricas, e outros ainda específicos para variáveis categóricas.

No melhoramento genético é possível trabalhar com mistura de variáveis já que existem informações quantitativas, qualitativas e moleculares. Todavia, muitos dos trabalhos não utilizam de todas as informações, excluindo da análise a possibilidade de se trabalhar de forma conjunta (BHERING et al.,2011). Um fator que contribui para isso é a dificuldade de se encontrar análises disponíveis em softwares que possibilitam ao pesquisador trabalhar com misturas.

Diante da baixa freqüência de estudos utilizando mistura de variáveis para se formar o agrupamento, resolveu-se trabalhar de forma conjunta, a fim de não ter nenhum tipo de limitação na análise e averiguar o quão importante é incluir ou não

variáveis dessa natureza à análise, gerando inclusive literatura para próximos trabalhos.

A proposta desse trabalho é avaliar dados fenotípicos advindos da simulação de cruzamentos de populações divergentes conhecidas a priori, para indicar dentre as combinações realizadas de distância e algoritmos de agrupamento, a mais eficiente para estudos futuros dentro da diversidade genética.

De forma geral, o trabalho busca avaliar agrupamentos em estudos de diversidade genética utilizando mistura de variáveis.

De forma específica, ele objetiva:

- Avaliar e comparar a eficiência dos agrupamentos obtidos das combinações existentes entre medidas de distância para variáveis numéricas (distância Euclidiana, Euclidiana ao quadrado, Euclidiana Média, Mahalanobis e Manhattan) com as medidas para variáveis categóricas (coeficiente de gower e coeficiente de concordância simples) com os métodos hierárquicos (Ligação Simples, Ligação Completa, UPGMA e Método da variância mínima de Ward) e os métodos não hierárquicos k-protótipos (HUANG, 1997) e k-médias;
- Identificar a combinação mais adequada para realizar estudos em diversidade genética quando se têm mistura de variáveis;
- Disponibilizar uma rotina em R para o algoritmo k-protótipos;
- Disponibilizar uma rotina em R para realizar agrupamentos com mistura de variáveis ou não utilizando métdos hierárquicos.

2. REVISÃO DE LITERATURA

1.1. Diversidade Genética

A diversidade genética é "qualquer medida quantitativa ou diferença genética, estando ao nível de seqüência ou nível de freqüência alélica, que é calculada entre indivíduos, populações ou espécies" (BEAUMONT et al., 1998; MOHAMMADI; PRASANNA, 2003).

O estudo da diversidade genética é de vital importância no melhoramento genético. Assim, a chance de se recuperar genótipos superiores é aumentada segundo Carvalho et al. (2003), quando se realiza cruzamentos entre genitores divergentes.

O objetivo da diversidade genética é elucidar relações genéticas, quantificar ou predizer o nível de variabilidade total existente e sua distribuição entre e/ou dentro de unidades taxonômicas, quer elas sejam indivíduos, acessos de bancos de germoplasma, linhagens, cultivares, populações ou espécies (BOLDT, 2011).

Existem dois meios de se estudar a diversidade genética. O primeiro é por técnicas biométricas relacionadas a quantificação da heterose e o segundo, por processos preditivos, o qual se destaca a análise de agrupamento e outras técnicas multivariadas.

Os autores Barbosa et al (2011) afirmaram que o uso da estatística multivariada também está presente para detectar a diversidade genética em estudos de espécies perenes e tem obtido sucesso.

Segundo ELIAS et al (2003), o uso de técnicas estatísticas multivariadas para se estimar a diversidade genética tem se tornado comum e é aplicado em várias culturas.

Para Viana et al (2003) trabalharam a diversidade genética entre genótipos comerciais de maracujazeiro-amarelo utilizando o complemento da similaridade de Jaccard nos marcadores avaliados e posteriormente o algoritmo de Ward, fazendo uso apenas de variáveis qualitativas; Coelho et al (2007) utilizaram a distância de Mahalanobis juntamente com o método agrupamento UPGMA em acessos de feijão, trabalhando exclusivamente com variáveis quantitativas.

Os autores Silveira et al (2009) definiram a diversidade genética entre cultivares de mandioca e espécies silvestres de *Manihot esculenta* através da análise

simultânea de variáveis morfológicas e posterior agrupamento utilizando variáveis quantitativas e qualitativas a partir do coeficiente de gower e o método UPGMA.

1.2. Análise de Agrupamento

A análise de agrupamento é amplamente utilizada em muitas áreas de pesquisa a fim de se reconhecer uma estrutura padrão de variabilidade entre os indivíduos ou objetos estudados, classificando-os em grupos homogêneos. A maioria dos trabalhos versa apenas sobre variáveis numéricas, excluindo da análise, as informações contidas nas variáveis categóricas.

De forma geral, pode-se obter uma matriz de dissimilaridade de três formas distintas: usando apenas variáveis numéricas; usando somente variáveis categóricas; e a utilização conjunta dessas variáveis, sendo que esse último procedimento pode ser segmentado em dois – utilizando um coeficiente que calcula a similaridade de uma só vez para essa mistura de variáveis (GOWER, 1971) ou de forma combinada, onde calcula-se a distância entre as observações usando métricas apropriadas para cada tipo específico de variável e no final, combina-se essas distâncias por algum peso γ pré-determinado (MINGOTI, 2005).

Dos procedimentos citados acima, a baixa existência de referências bibliográficas aliada aos poucos métodos de agrupamento implementados nos softwares estatísticos e de *data-mining*, faz com que a análise conjunta das variáveis seja dificultada e esquecida.

1.2.1. Medidas de Similaridade e Dissimilaridade

Seja uma amostra aleatória de *n* elementos amostrais, com *p* variáveis aleatórias. O vetor de medidas $\mathbf{X}_j$ é:

$$\mathbf{X}_j = [X_{1j} \; X_{2j} \; \ldots \; X_{pj}]', \; j = 1, 2, \ldots, n$$

em que X_{ij} representa o valor observado da variável *i* medida no objeto *j*.

Para realizar a análise de agrupamento é necessário primeiramente definir qual medida de similaridade ou dissimilaridade será usada. Dentre as muitas

existentes, será apresentado nessa seção algumas das medidas avaliadas para variáveis contínuas e posteriormente, medidas próprias para variáveis categóricas, além é claro de medidas específicas para misturas de variáveis.

1.2.1.1. Dissimilaridade em Variáveis Contínuas

A seguir serão apresentadas as métricas utilizadas nessa dissertação.

1.2.1.1.1. Distância Euclidiana

A distância Euclidiana entre dois elementos X_l e X_k, $l \neq k$ é a medida de dissimilaridade mais utilizada e conhecida e é dada por:

$$d(X_l,X_k) = [(X_l - X_k)^T (X_l - X_k)]^{1/2} = [\sum_{i=1}^{p} (X_{il} - X_{ik})^2]^{1/2}$$

O quadrado da distância Euclidiana é definido como:

$$d(X_l,X_k) = [(X_l - X_k)^T (X_l - X_k)] = \sum_{i=1}^{p} (X_{il} - X_{ik})^2$$

1.2.1.1.2. Distância de Mahalanobis

Na distância de Mahalanobis é acrescentada a informação sobre possíveis diferenças de variâncias e as relações lineares entre as variáveis, dadas pela inversa da matriz de variâncias e covariâncias S_{pxp}^{-1}. Dessa forma, a medida de Mahalanobis é:

$$\mathbf{d}(X_l,X_k) = [(X_l - X_k)^T S_{pxp}^{-1} (X_l - X_k)]^{1/2}$$

1.2.1.1.3. Distância Euclidiana Média

A diferença da distância Euclidiana Média para a de Mahalanobis está na matriz de variâncias e covariâncias que é substituída pela matriz diagonal do inverso do número de variáveis (diag (1/p))

$$d(X_l,X_k) = [(X_l - X_k)^T \, diag(1/p)(X_l - X_k)]^{1/2}$$

Como a distância Euclidiana cresce com o aumento do número de variáveis, essa distância consegue eliminar o efeito do número de variáveis ao utilizar a matriz $diag(1/p)$.

1.2.1.1.4. Distância de Minkowsky

A distância de Minkowsky entre dois elementos X_l e X_k, $l \neq k$ é escrita como:

$$d(X_l,\mathbf{X}_k) = [\sum_{i=1}^{p} w_i \left| X_{il} - X_{ik} \right|]^{1/\lambda}$$

em que w_i é o peso de ponderação para a variável i. Para λ=1tem-se a distância de Manhattan e, se for adotado λ=2, tem-se a distância Euclidiana.

Segundo MINGOTI(2005), essa distância é menos afetada pela presença de outliers do que se comparado à distância Euclidiana.

1.2.1.2. Métrica em Variáveis Categóricas

A seguir serão apresentadas as medidas de similaridade utilizadas para as variáveis categóricas.

1.2.1.2.1. Coeficiente de Concordância simples

O coeficiente de concordância simples mede a similaridade (parecença) entre dois indivíduos X_l e X_k, $l \neq k$, e este é calculado como a proporção de categorias similares entre tais elementos. Veja o exemplo a seguir:

Tabela 1 – Observações de dois elementos amostrais segundo três atributos

Variável	V1	V2	V3
Elemento 1	0	1	1
Elemento 2	0	1	0

O coeficiente de concordância simples entre os elementos 1 e 2 é

$$S(1,2) = 2/3 = 0,67$$

pois existem dois pares concordantes ((1,1) e (0,0)) em um total de 3 pares.

1.2.1.2.2. Coeficiente de Jaccard

O coeficiente de Jaccard é similar ao de concordância simples, porém é aplicado somente a variáveis dicotômicas. Mede a parecença entre dois indivíduos calculando a proporção de pares do tipo (1 1) sobre o número total de pares possíveis de serem comparados, ou seja, excluindo os pares do tipo (0 0) já que ambos não possuem a característica de interesse.

Assim, o coeficiente de Jaccard para esse mesmo exemplo é:

$$S(1,2) = 1/3 = 0,33$$

1.2.1.3. Métrica para mistura de variáveis

A seguir será apresentado o coeficiente de Gower (1971), que trabalha tanto com variáveis numéricas quanto categóricas.

1.2.1.3.1. Coeficiente de Gower (1971)

O coeficiente de Gower concentra uma medida de similaridade específica para variáveis contínuas, multicategóricas e binárias.

A similaridade entre os indivíduos X_l e X_k, $l \neq k$, é expresso na fórmula a seguir:

$$S_{lk} = \frac{\sum_{i=1}^{p} s_{lki}\delta_{lki}}{\sum_{i=1}^{p} \delta_{lki}}$$

em que s_{lki} é a similaridade entre os indivíduos X_l e X_k na variável *i* e δ_{lki} é uma função indicadora que indica se é possível comparar tais indivíduos para determinada variável *i*.

No caso em que se têm variáveis contínuas, a similaridade s_{lki} é expressa como

$$s_{lki} = 1 - \frac{|x_l - x_k|}{R_i}$$

em que x_l e x_k são os valores assumidos pelos indivíduos X_l e X_k, $l \neq k$ para a variável *i*.

R_i é a amplitude da variável *i*, fazendo com que o coeficiente s_{lki} esteja entre 0 e 1.

Para o caso das variáveis categóricas, a similaridade é dada por uma função que recebe 1 caso dois indivíduos sejam similares e 0, caso contrário.
O coeficiente de Gower contempla, inclusive, a comparação de elementos amostrais com informações incompletas (MINGOTI, 2005).

1.2.1.3.2. Distâncias Combinadas

Uma outra forma de estudar as variáveis conjuntamente é combinando medidas específicas para variáveis contínuas e categóricas através de um peso (γ) pré-determinado.

Veja, por exemplo, que é possível combinar qualquer medida de distância específica para variáveis quantitativas com qualquer medida específica para variáveis qualitativas.

$$\mathrm{d}(\mathrm{X}_l, \mathbf{X}_k) = \left[\sum_{i=1}^{pcont} \left| X_{il} - X_{ik} \right|^2 \right] + \gamma \sum_{i=1}^{pcat} \delta_{X_{il}, X_{ik}}$$

o primeiro termo da expressão corresponde a distância Euclidiana ao quadrado e o segundo termo, a uma função dicotômica que recebe 1 caso os indivíduos X_l e X_k não pertençam a mesma categoria e 0, caso pertençam.

Vale deixar claro que poderia ter sido usada outra distância específica para variáveis quantitativas como, por exemplo, a distância de Mahalanobis ou outra qualquer.

1.2.2. Técnicas de Agrupamento

A seguir encontram-se os métodos de agrupamentos utilizados nesse trabalho. Dentre os algoritmos que seguem, apenas o k-Protótipos e o k-Médias são da classe dos não-hierárquicos. É importante lembrar que em todos os métodos hierárquicos, a ideia é agrupar os indivíduos mais similares e que estes se diferenciam na forma como atualizam a matriz de distâncias D_{nxn} (construída a partir das métricas mostradas anteriormente).

1.2.2.1. Método de Ligação Simples (Vizinho mais próximo)

Nesse método de agrupamento hierárquico, a distância entre dois grupos C_1 e C_2 é definido pela distância mínina existente entre os indivíduos pertencentes a tais grupos.

$$d(C_1, C_2) = \min(d(X_l, X_K, l \neq k))$$

Primeiramente, constrói-se a matriz de distância $\boldsymbol{D_{nxn}}$ e então vão sendo formados os grupos a partir das menores distâncias. A cada vez que um elemento é agrupado a um determinado grupo, a matriz de distância é atualizada baseando-se na menor distância entre os indivíduos de dois grupos, caracterizando assim o método.

1.2.2.2. Método de Ligação Completa (Vizinho mais distante)

Ao contrário do método exposto anteriormente, o método do vizinho mais distante atualiza a matriz de distâncias calculando a distância máxima existente entre os indivíduos de dois grupos.

$$d(C_1, C_2) = \max(d(X_l, X_K, l \neq k))$$

1.2.2.3. Método UPGMA (Unweighted Pair-Group Method using Arithmetic Averages)

Nesse método, a matriz de distâncias é atualizada calculando-se a média das distâncias entre os indivíduos de dois grupos. Assim, se C_1 tem n_1 indivíduos e C_2 tem n_2 indivíduos, a distância entre eles será definida por

$$d(C_1, C_2) = \sum_{l \in C_1} \sum_{k \in C_2} \left(\frac{1}{n_1 n_2} \right) d(X_l, X_k)$$

Esse método visa trabalhar com médias ao invés de valores extremos.

1.2.2.4. Método de Ward

O método de Ward (WARD, 1963) ou de variância mínima consiste em formar grupos a partir de pares que proporcionem a menor soma de quadrados.

Cada elemento é considerado um conglomerado e então, calcula-se a soma de quadrados dentro de cada conglomerado. Esta soma é o quadrado da distância Euclidiana de cada elemento pertencente ao conglomerado em relação ao correspondente vetor de médias do conglomerado

$$SS_i = \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i.} \right)^T \left(X_{ij} - \bar{X}_{i.} \right)$$

em que n_i é o número de elementos do conglomerado C_i quando se está no passo k do processo de agrupamento; X_{ij} é o vetor de observações do *j*-ésimo elemento pertencente ao *i*-ésimo conglomerado; $\bar{X}_{i.}$ é o vetor de médias do conglomerado C_i e SS_i é a soma de quadrados referente a tal conglomerado (MINGOTI, 2005).

Posteriormente, calcula-se a soma de quadrados entre dois conglomerados C_l e C_i que é dado por:

$$d(C_l, C_i) = \left[\frac{n_l n_i}{n_l + n_i} \right] \left(\bar{X}_{l.} - \bar{X}_{i.} \right)^T \left(\bar{X}_{l.} - \bar{X}_{i.} \right)$$

em que $\left[\frac{n_l n_i}{n_l + n_i} \right]$ é um fator de ponderação para quando os conglomerados tiverem tamanhos diferentes (MINGOTI, 2005).

A cada passo do algoritmo, os dois conglomerados que minimizam tal distância são combinados.

1.2.2.5. Algoritmo K-médias

É um algoritmo da classe dos não hierárquicos que tem por objetivo minimizar a distância dos elementos a um conjunto de centróides de forma iterativa.

Seu parâmetro é o número k de clusters que é definido a priori pelo pesquisador. Outra questão a ser mencionada é que é necessário entrar com "sementes iniciais" para inicializar o algoritmo e isso pode ser feito de várias formas, como cita Mingoti (2005):

- Método da escolha pré-fixada;
- Escolha aleatória;
- Escolha via variável aleatória;
- Valores discrepantes;
- K primeiros valores do banco de dados.

Os passos do algoritmo são:

1. Escolhe-se k centróides, chamados de "sementes", para iniciar o processo;
2. Através da distância Euclidiana, cada elemento é comparado com cada centróide inicial (semente). Daí, o elemento é alocado ao grupo cuja distância é a menor;
3. Após o passo 2, calcula-se os valores dos centróides para cada novo grupo formado e então, repete-se o passo 2;
4. Os passos 2 e 3 são repetidos até que nenhuma realocação de elementos seja necessária.

1.2.2.6. Algoritmo k-Protótipos

O algoritmo de k-protótipos, proposto por Huang (1997, 1998) é uma extensão do conhecido algoritmo não-hierárquico k-Médias. O k-protótipos remove a

limitação do k-Médias, ao trabalhar não somente com variáveis contínuas, mas também com variáveis categóricas.

Quando é aplicado somente a variáveis contínuas, o k-protótipos possui o mesmo comportamento que o k-Médias, e ainda segundo Huang (1997), é eficiente na análise de base de dados grande e complexa.

Muito utilizado em *Data Mining*, esse algoritmo comporta uma medida de distância combinada para poder trabalhar com essa mistura de variáveis,

$$d(X_l, X_k) = \left[\sum_{i=1}^{pcont} X_{il} - X_{ik}^{\ 2} \right] + \gamma \sum_{i=1}^{pcat} \delta_{X_{il}, X_{ik}}$$

em que a medida de distância para variáveis contínuas é o quadrado da distância Euclidiana e para as variáveis categóricas, a medida é uma função indicadora que recebe 1 caso os indivíduos X_l e X_k não pertençam a mesma categoria e 0, caso pertençam. Há ainda o uso de um peso γ para a medida de dissimilaridade das variáveis categóricas que, segundo proposta de Huang (1997), pode ser a média dos desvios padrão de todas as variáveis contínuas, ou seja, como a maior dispersão de tais variáveis ocasiona uma super medida de distância, o uso das médias dos desvios padrão é uma forma de tentar equilibrar o grau de importância em termos de tipo de variável.

Além dessa medida sugerida pelo autor, pode-se construir outras diversas medidas combinadas, como foi feito nesse trabalho.

O k-protótipos possui os seguintes passos:

1. Escrever os protótipos (sementes) iniciais;
2. Alocar cada indivíduo do banco de dados no protótipo de menor distância, de acordo com a medida que foi apresentada;
3. Atualizar os protótipos iniciais com médias e modas, de acordo com o tipo de variável (quantitativa ou qualitativa);
4. Realocar os indivíduos nos protótipos até que não haja nenhuma mudança.

2. MATERIAL E MÉTODOS

Com o propósito de avaliar as medidas e os métodos de agrupamentos em populações bem definidas geneticamente para avaliar qual o algoritmo de agrupamento mais adequado aliado a uma dada medida de dissimilaridade no tratamento desse tipo de dados, foram gerados no software GENES, versão 2011.9.0, 13 populações com 200 genótipos e dez características fenotípicas quantitativas, formando uma matriz de 2600 linhas por 10 colunas.

Na projeção das dez populações (Figura 1), observa-se que as populações 1 e 9 estão mais dispersas pela distância Euclidiana.

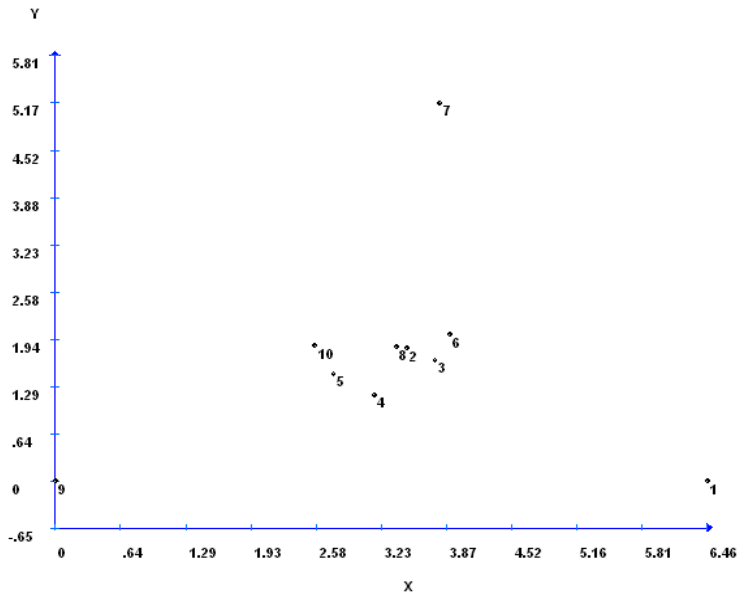

Figura 1 - Projeção 2D das 10 populações

Como 1 e 9 foram as populações mais divergentes derivadas dessa primeira simulação, estas serão chamadas de pais 1 e 2 (P_1 e P_2) e serão utilizadas para cruzamento entre si, e em seguida, para realizar retrocruzamentos.

A Tabela 2 mostra o percentual dos genes relacionadas aos genitores P_1 e P_2 para F_1 e demais retrocruzamentos.

Chamou-se F_1 de população 3, o retrocruzamento do P_1 com F_1 (RC_{11}) de população 4, o retrocruzamento do P_1 com RC_{11} de população 5 e assim por diante, até se realizar cinco retrocruzamentos para cada genitor.

Tabela 2 – Percentual dos genitores 1 e 2 nos cruzamentos e retrocruzamentos

Cruzamentos e Retrocruzamentos	Genitores	
	P_1	P_2
F_1 (3)	50%	50%
RC_{11} (4)	75%	25%
RC_{12} (5)	87,5%	12,5%
RC_{13} (6)	93,75%	6,25%
RC_{14} (7)	96,875%	3,125%
RC_{15} (8)	98,4375%	1,5625%
RC_{21} (9)	25%	75%
RC_{22} (10)	12,5%	87,5%
RC_{23} (11)	6,25%	93,75%
RC_{24} (12)	3,125%	96,875%
RC_{25} (13)	1,5625%	98,4375%

A justificativa para se realizar cinco retrocruzamentos é que dessa forma as populações resultantes (RC_{15} e RC_{25}) se aproximam com 98,4375% da estrutura genética dos pais teoricamente, que são os mais próximos geneticamente de P_1 e P_2, respectivamente.

Dessa forma, poderia-se condensar as 13 populações em apenas três conglomerados como foi feito abaixo:

- Grupo 1: P_1 (1), RC_{11} (4) , RC_{21} (5) ,RC_{31} (6) ,RC_{41} (7) ,RC_{51} (8) ;
- Grupo 2: P_2 (2) , RC_{12} (9), RC_{22} (10), RC_{32} (11), RC_{42} (12), RC_{52} (13);
- Grupo 3: F_1 (3) .

A Figura 2 mostra como está estruturado tanto o cruzamento dos genitores, quanto os retrocruzamentos.

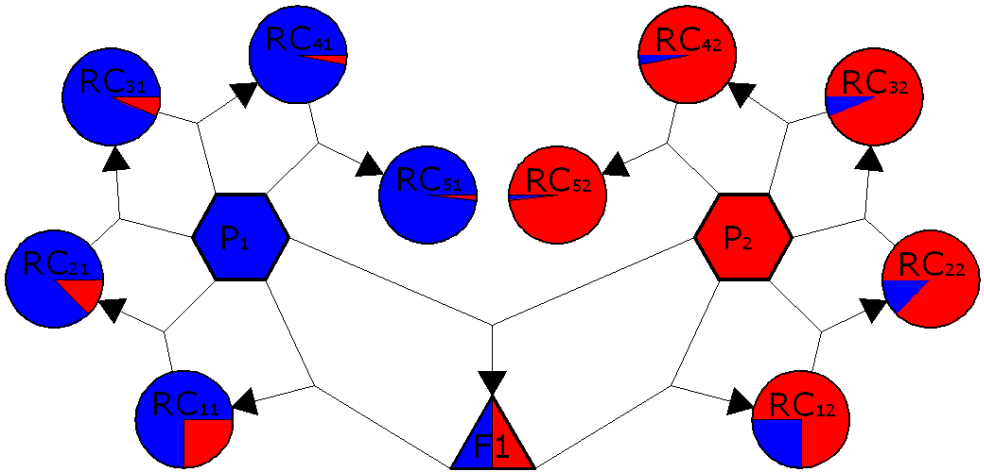

Figura 2 – Fluxograma dos cruzamentos e retrocruzamentos

Ainda discutindo a simulação, foram geradas dez variáveis quantitativas em que a herdabilidade variou de 25 à 70% entre as variáveis, pois pretendia-se avaliar o quão influente poderia ser o meio ao realizar um agrupamento. Assim, procurou-se simular variáveis com diferentes herdabilidades.

Segundo CARDELLINO e OSÓRIO (1999), a herdabilidade varia de 0 a 100%, sendo que valores abaixos de 20% são considerados baixos, de 20% a 40% moderados e acima de 40%, valores altos.

OLIVEIRA et al., (2007) citaram que quanto mais alta é a herdabilidade de uma característica, melhor é a predição do valor genético pelo desempenho individual e mais rápida é a resposta à seleção para essa característica.

Outra questão a ser mencionada é que, como as dez variáveis geradas possuem médias distintas, padronizou-se as variáveis dividindo-as pelos seus respectivos desvios-padrão para não priorizar determinada variável na realização do método de agrupamento. Assim, todas as variáveis passaram a ter desvio-padrão igual a 1.

Como esse estudo está avaliando métodos de agrupamento e medidas de dissimilaridade para mistura de variáveis (quantitativas e qualitativas) e as variáveis simuladas são de natureza quantitativa, foi estabelecido que a categorização das variáveis fosse realizada conforme tabela abaixo.

Tabela 3 – Categorização das variáveis quantitativas segundo o método dos percentis

Valores menores ou iguais ao percentil 50	0
Valores maiores que o percentil 50	1

Pela tabela observa-se que as variáveis categóricas serão representadas apenas pelas binárias, excluindo do processo variáveis multicategóricas.

Outro ponto importante é que foram feitos vários cenários distintos. Primeiramente foram avaliados agrupamentos trabalhando somente com variáveis quantitativas. Depois, foi incluído uma variável binária de cada vez na análise, de acordo com a Tabela 4.

Tabela 4 – Cenários de avaliação do número de variáveis quantitativas e qualitativas

Cenários	Variáveis Quantitativas	Variáveis Binárias
1	10	0
2	9	0
3	8	0
4	7	0
5	6	0
6	5	0
7	4	0
8	3	0
9	2	0
10	9	1
11	8	2
12	7	3
13	6	4
14	5	5
15	4	6
16	3	7
17	2	8

Ao utilizar mistura de variáveis, sempre a última variável quantitativa será categorizada. Por exemplo, realizado todos os agrupamentos utilizando só variáveis numéricas, o próximo cenário, que é o décimo, utilizará de nove variáveis quantitativas e uma binária, e a categorização se dará na variável dez que possui o maior percentual de herdabilidade (70%); no cenário seguinte, onde são oito variáveis quantitativas e duas binárias, além da categorização da variável dez, será categorizado a variável nove, que possui 65% de herdabilidade, e assim até chegar ao último cenário que possui duas variáveis quantitativas e oito binárias.

Como forma de atribuir peso às variáveis binárias, o parâmetro γ, citado no capítulo 2, terá quatro valores que são de acordo com o número de desvios-padrão das variáveis quantitativas, sugeridos por Huang (1997).

Tabela 5 - Valores para o parâmetro gamma associado às variáveis categóricas

Desvios-Padrão	γ
0,5	0,5
1	1
2	2
3	3

As medidas de dissimilaridade utilizadas foram:

- Distância Euclidiana;
- Distancia Euclidiana Combinada;
- Distância Euclidiana ao quadrado;
- Distancia Euclidiana ao quadrado Combinada;
- Distância Euclidiana Média;
- Distancia Euclidiana Média Combinada;
- Distância Mahalanobis;
- Distancia Mahalanobis Combinada;
- Distancia Manhattan;
- Distancia Manhattan Combinada;
- Distancia de Gower.

Em que as distâncias combinadas são simplesmente a junção da medida de dissimilaridade para variáveis contínuas com uma função dicotômica do tipo 0 ou 1, proposta por Huang (1997).

Os métodos de agrupamento hierárquicos são:

- Ligação Simples (vizinho mais próximo);
- Ligação Completa (vizinho mais distante);
- UPGMA;
- Ward.

Os métodos de agrupamento não hierárquicos são:

- K-protótipos;
- K-médias.

Toda a análise de agrupamento foi realizada no software R, versão 2.12.2 através do pacote *StatMatch* que calcula a distância de Gower pela função *gower.dist*, e também pelas funções *dist*, *mahalanobis.dist*, *hclust* e *kmeans*. Essas funções foram utilizadas para realizar os agrupamentos hierárquicos e também o algoritmo k-médias.

Como o algoritmo k-protótipos não estava implementado em nenhum software, implementou-se o em R.

A seguir encontra-se um pequeno exemplo motivacional sobre o algoritmo k-protótipos.

Na tabela encontra-se duas variáveis numéricas V1 e V2 e uma variável binária V3. Pretende-se formar dois grupos utilizando o algoritmo k-protótipos.

Tabela 6 – Observações de elementos amostrais segundo três atributos

Variável	V1	V2	V3
Elemento 1	20	7	1
Elemento 2	39	2	0
Elemento 3	18	10	1

Como pretende-se formar dois grupos, é necessário 2 protótipos iniciais. Cada um dos protótipos contém uma média para cada variável numérica e uma moda para a variável binária.

Seja os protótipos iniciais e considere γ=1.

- Protótipo 1 = (18;8;1);
- Protótipo 2 = (17;9;0).

As distâncias do Elemento 1 aos Protótipos 1 e 2 são:

$$\text{d(Elemento 1, Protótipo 1)} = (20-18)^2 + (7-8)^2 + 1x0 = 5$$

$$\text{d(Elemento 1, Protótipo 2)} = (20-17)^2 + (7-9)^2 + 1x0 = 13$$

Assim verifica-se que

$$\text{d(Elemento 1, Protótipo 1)} < \text{d(Elemento 1, Protótipo 2)}$$

e a observação 1 é alocada ao grupo 1, que refere-se ao Protótipo 1.

O mesmo procedimento é feito para os elementos 2 e 3, os quais são alocados nos grupos 1 e 2, respectivamente.

Após a primeira iteração, os protótipos são atualizados calculando médias e modas dos elementos amostrais alocados nesses grupos. Como o grupo 1,

representado por G_1 possui os elementos 1 e 2, e como o grupo 2, representado por G_2, possui o elemento 3, a atualização dos protótipos é:

- Protótipo 1 = (29,5;4,5;0);
- Protótipo 2 = (18;10;1).

Na segunda iteração, as distâncias são:

d(Elemento 1, Protótipo 1) $= 97,5$

d(Elemento 1, Protótipo 2) $= 13$

d(Elemento 2, Protótipo 1) $= 96,5$

d(Elemento 2, Protótipo 2) $= 506$

d(Elemento 3, Protótipo 1) $= 163,5$

d(Elemento 3, Protótipo 2) $= 0$

Após a segunda iteração, os protótipos 1 e 2 devem ser atualizados calculando médias e modas dos elementos amostrais alocados nesses grupos. Agora, G_1=(Elemento 2) e G_2=(Elemento 1, Elemento 3) e a atualização dos protótipos será

- Protótipo 1 = (39;2;0);
- Protótipo 2 = (19;8,5;1).

Repetindo os cálculos com os novos protótipos, nenhum elemento amostral muda de grupo, ou seja, o algoritmo é finalizado.

Assim, os grupos são G_1=(Elemento 2) e G_2=(Elemento 1, Elemento 3).

3. RESULTADOS E DISCUSSÃO

Foi avaliado o desempenho das combinações realizadas entre as distâncias e os métodos de agrupamento hierárquicos e não hierárquicos em cenários que continham número de variáveis quantitativas e binárias diferentes, além do "peso" γ associado às variáveis binárias, a fim de se determinar qual a combinação mais eficiente, ou seja, a que possuía a maior taxa de acerto quanto à alocação dos genótipos.

Cada uma das 13 populações possui 200 indivíduos, as quais poderiam ser condensadas em três grupos a se saber:

- Grupo 1: P_1 (1), RC_{11} (4) , RC_{21} (5) ,RC_{31} (6) ,RC_{41} (7) ,RC_{51} (8) ;
- Grupo 2: P_2 (2) , RC_{12} (9), RC_{22} (10), RC_{32} (11), RC_{42} (12), RC_{52} (13);
- Grupo 3: F_1 (3) .

e como forma de motivação e melhor entendimento dos resultados, apresenta-se o dendograma construído a partir das médias dessas populações para as dez variáveis quantitativas para evidenciar não só a eficiência do método utilizado e sua distância, como também a confirmação que a simulação foi bem estruturada.

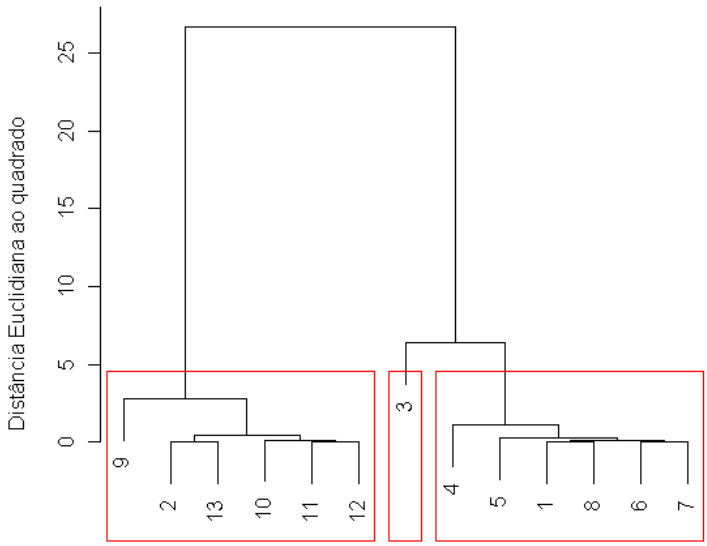

Figura 3 – Dendograma das médias das 13 populações utilizando a distância Euclidiana ao quadrado com o método de Ward

Pela Figura 3 observa-se claramente as populações muito bem definidas em seus grupos. Os métodos do vizinho mais próximo e UPGMA também tiveram resultados semelhantes a esse para a mesma distância empregada. Porém, como não conhecemos na prática a população a que pertence os genótipos, seria impossível trabalhar apenas com as médias das observações para gerar o agrupamento. Por isso esse trabalho visa agrupar indivíduos e não médias de populações.

A população F_1 (3), resultante do cruzamento dos P_1 e P_2, que possui 50% das características de cada um de seus pais, foi excluída da análise já que estava gerando confundimento.

As tabelas a seguir mostram alguns dos métodos e distâncias aplicados em dois cenários distintos quanto à herdabilidade, que resultaram em grupos onde não foi possível identificar o desempenho deles devido ao confundimento gerado por F_1.

É preciso salientar que o *label* dos grupos é apenas uma forma de identificar as populações no agrupamento. Em negrito encontra-se o maior número de indivíduos classificados nos grupos 1, 2 e 3 para as três primeiras populações.

Tabela 7 - Cinco características de menor herdabilidade – Distância Mahalanobis e Método Vizinho mais distante

População	Grupo 1	Grupo 2	Grupo 3
1	**143**	50	7
2	**147**	11	42
3	**135**	57	8
4	136	55	9
5	135	51	14
6	138	47	15
7	143	45	12
8	144	49	7
9	138	34	28
10	142	16	42
11	145	18	37
12	143	19	38
13	142	19	39

Tabela 8 - Cinco características de menor herdabilidade – Distância Mahalanobis e Método Vizinho mais distante

População	Grupo 1	Grupo 2	Grupo 3
1	**106**	78	16
2	84	26	**90**
3	**124**	61	15
4	109	74	17
5	116	70	14
6	117	71	12
7	107	79	14
8	113	77	10
9	104	43	53
10	88	40	72
11	92	40	68
12	89	31	80
13	91	37	72

Tabela 9 - Cinco características de menor herdabilidade – Distância Euclidiana e Método Vizinho mais distante

População	Grupo 1	Grupo 2	Grupo 3
1	**95**	70	35
2	18	73	**109**
3	**149**	36	15
4	104	63	33
5	109	58	33
6	86	83	31
7	91	69	40
8	89	77	34
9	76	66	58
10	43	74	83
11	38	64	98
12	29	71	100
13	24	75	101

Tabela 10 - Cinco características de maior herdabilidade – Distância Mahalanobis e Método Vizinho mais distante

População	Grupo 1	Grupo 2	Grupo 3
1	**118**	62	20
2	**106**	58	36
3	40	**140**	20
4	70	100	30
5	80	89	31
6	108	65	27
7	93	85	22
8	95	73	32
9	61	111	28
10	75	88	37
11	88	79	33
12	91	75	34
13	93	69	38

Por essas tabelas, observa-se que não foi possível distinguir as populações do estudo, já que, na última tabela, por exemplo, 118 genótipos da população 1 foram classificados no grupo 1 junto com 106 indivíduos da população 2. Ora, a população 1 e 2 são as mais divergentes e teriam que ser alocadas em grupos diferentes e isso não ocorreu. Já a população 3 teve 140 genótipos no grupo 2.

Dessa forma, esses resultados conduziram a excluir a população 3 do estudo de forma a viabilizar a mensuração do desempenho dos métodos.

Antes de excluir de fato a população 3, foi feito um levantamento a cerca da taxa de acerto levando em consideração todos os indivíduos pertencentes as 3 primeiras populações para todas as 10 características, excluindo assim todos os retrocruzamentos. Ou seja, tem-se a seguir alguns dos resultados do agrupamento gerado para 600 genótipos.

Tabela 11 - Resultado do agrupamento para o Genitor 1 (P_1), Genitor 2 (P_2) e população F1 – Distância Euclidiana ao quadrado e Método de Ward

População	Grupo 1	Grupo 2	Grupo 3
1	**182**	16	2
2	28	**170**	2
3	98	2	**100**

Tabela 12 - Resultado do agrupamento para o Genitor 1 (P_1), Genitor 2 (P_2) e

população F1 – Distância Euclidiana e Método de Ward

População	Grupo 1	Grupo 2	Grupo 3
1	**180**	20	0
2	27	**171**	2
3	101	3	**96**

Para esses duas tabelas apresentadas, observa-se com clareza a distinção gerada no agrupamento pelo método de Ward e distâncias Euclidiana ao quadrado e Euclidiana, respectivamente, onde o primeiro teve um índice de acerto de 75,33% e o segundo, 74,50%.

Contudo, outros resultados mostraram a ineficiência do método do vizinho mais próximo junto com a distância de Mahalanobis. Pela Tabela 13 conclui-se que não foi possível discriminar as populações 1, 2 e 3, resultando em apenas 40,83% de acerto.

Tabela 13 - Resultado do agrupamento para o Genitor 1 (P_1), Genitor 2 (P_2) e população F1 – Distância Euclidiana e Método de Ward

População	Grupo 1	Grupo 2	Grupo 3
1	**118**	33	49
2	93	15	**92**
3	85	**35**	80

Mais uma vez é importante ressaltar que o nome dos grupos é simplesmente uma forma de distinguir os indivíduos das populações.

Por tais motivos mencionados, esse trabalho irá focar no desempenho dos métodos excluindo do estudo a população 3.

3.1. Métodos Hierárquicos

Para se realizar um agrupamento pelo método hierárquico deve-se, primeiramente, escolher a medida de distância apropriada e então, aplicar algum algoritmo hierárquico.

Foram utilizados quatro algoritmos hierárquicos nesse estudo:

- Ligação Simples (vizinho mais próximo);
- Ligação Completa (vizinho mais distante);
- UPGMA;
- Ward.

Desses, os métodos do vizinho mais próximo e o UPGMA se mostraram ineficazes ao realizar o agrupamento já que não conseguiram fazer distinção dos 2600 genótipos avaliados. Para MINGOTI (2005), o método de Ligação Simples (Vizinho mais próximo) não é capaz de delinear grupos pouco separados.

Mesmo variando os cenários quanto ao número de variáveis quantitativas ou binárias e também quanto ao parâmetro γ, esses métodos se mostraram ineficientes já que não separaram os indivíduos, resultando num único conglomerado.
O dendograma a seguir mostra a distância Euclidiana utilizada junto ao método do vizinho mais próximo. Observe que os genótipos não foram distinguidos.

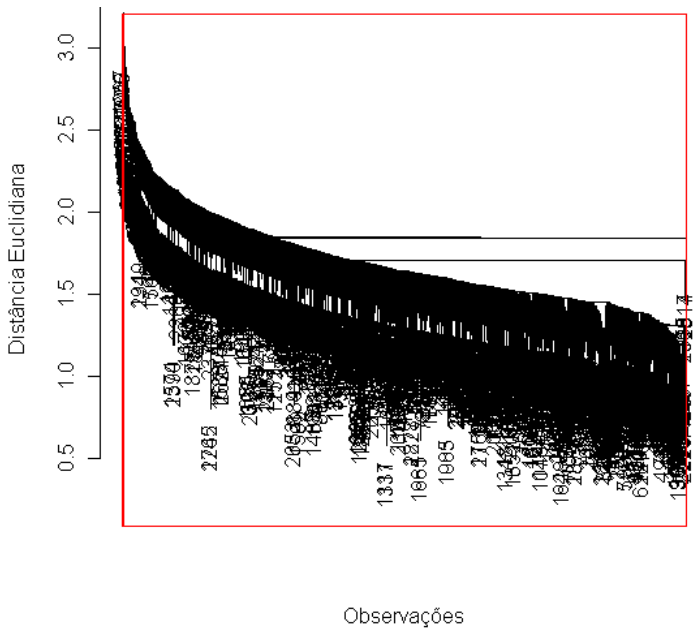

Figura 4 – Dendograma de 2600 genótipos utilizando a distância Euclidiana com o método hierárquico Ligação Simples (Vizinho mais próximo)

Como conseqüência desses resultados, será apresentado o desempenho apenas dos métodos Ligação Completa (vizinho mais distante) e Ward.

É interessante ressaltar que quando utilizou-se das médias das populações para fazer o agrupamento, os métodos Ligação Simples (vizinho mais próximo) e

UPGMA se mostraram tão eficientes quanto o de Ward, enquanto o Ligação Completa agrupou um indivíduo errado (população 9).

Porém, ao se trabalhar com todos os genótipos, apenas dois dos quatro métodos são propícios ao tratamento dos dados.

3.2. Métodos não hierárquicos

Os métodos não hierárquicos utilizados nessa dissertação foram o método das k-médias (*k-means*) e o k-protótipos (HUANG, 1997), derivado do primeiro.

Como já foi apresentado todos os métodos de agrupamento utilizados juntamente com as distâncias no capítulo 3 (Material e Métodos), encontra-se na seção seguinte os resultados.

Primeiramente será apresentado o agrupamento resultante apenas da avaliação de variáveis numéricas.

Posteriormente, o desempenho dos métodos envolvendo mistura de variáveis será mostrado.

3.3. Análise de agrupamento usando somente variáveis quantitativas

A tabela a seguir mostra o desempenho dos métodos hierárquicos (Ligação Completa e Ward) e do não hierárquico k-médias quando considerou-se todos os cenários utilizando somente variáveis quantitativas.

É relevante informar que as sementes iniciais utilizadas no algoritmo de k-médias foram as duas primeiras observações do banco de dados. Outra informação importante é que quando não foi possível mensurar o desempenho de algum algoritmo combinado a uma determinada distância, utilizou-se o símbolo "-". Os desempenhos que foram iguais ou superiores a 80% foram destacados em negrito para chamar a atenção do leitor.

Pelo exposto, observa-se que o método de Ward foi superior ao método de Ligação Completa para todas as distâncias utilizadas ao avaliar todas as dez variáveis. O algoritmo não hierárquico k-médias também obteve notório desempenho ao agrupar cerca de 81,29% dos genótipos corretamente.

Quando o método de Ligação Completa foi utilizado após se construir a matriz de distâncias pelas medidas de dissimilaridade Euclidiana ao quadrado,

Euclidiana Média e Manhattan este não obteve resultados coerentes, impossibilitando contabilizar o percentual de acerto pois não distinguiu as populações 1 e 2. Assim utilizou-se do símbolo "-" para representar os agrupamentos onde não foi possível fazer tal distinção das populações.

Ao se trabalhar com nove variáveis quantitativas, excluindo a variável com maior índice de herdabilidade (70%), supondo que esta fosse categórica, ou seja, em um cenário onde o pesquisador possui a variável categórica, mas não a utiliza pode-se observar que os maiores desempenhos de qualidade do agrupamento são utilizando o método de agrupamento de Ward, com exceção da medida de dissimilaridade Euclidiana Média que junto ao método Ligação Completa teve bom resultado. As distâncias de Manhattan e Gower tiveram as maiores taxas de acerto para o método de Ward. O método das k-médias obteve resultado superior a todos os outros.

Tabela 14 – Taxa de desempenho utilizando todos os cenários somente avaliando variáveis quantitativas

Métodos	Número de variáveis quantitativas utilizadas								
	10	9	8	7	6	5	4	3	2
Euclidiana – Ligação Completa	70,08%	75,25%	-	68,92%	67,79%	66,17%	69,25%	65,38%	54,38%
Euclidiana – Ward	**81,29%**	74,58%	**81,00%**	**81,00%**	71,21%	65,92%	74,42%	-	-
Euclidiana ao quadrado – Ligação Completa	-	75,25%	-	68,92%	67,79%	66,17%	69,25%	65,38%	54,38%
Euclidiana ao quadrado – Ward	**81,63%**	79,83%	79,63%	76,83%	62,13%	70,29%	74,17%	71,54%	69,38%
Euclidiana Média – Ligação Completa	-	79,83%	-	61,83%	57,04%	69,21%	-	63,54%	54,38%
Euclidiana Média – Ward	78,00%	77,46%	**82,92%**	72,00%	71,75%	66,58%	73,67%	-	-
Mahalanobis – Ligação Completa	71,83%	-	-	-	-	-	-	64,33%	-
Mahalanobis – Ward	76,13%	67,54%	71,96%	71,29%	68,88%	62,08%	59,75%	64,96%	60,67%
Manhattan – Ligação Completa	-	74,96%	74,00%	-	-	-	-	-	64,29%
Manhattan – Ward	76,13%	**80,54%**	77,88%	77,00%	69,75%	58,50%	74,96%	67,96%	62,79%
Gower - Ligação Completa	69,50%	76,04%	73,17%	64,63%	67,29%	65,46%	-	66,33%	67,63%
Gower – Ward	**80,67%**	**80,17%**	77,63%	**80,17%**	74,29%	-	71,33%	72,25%	70,13%
k-médias	**81,29%**	**81,54%**	**83,00%**	**81,67%**	74,04%	72,63%	77,08%	76,29%	67,08%

Ao utilizar oito variáveis quantitativas (excluiu-se as variáveis com herdabilidade 70% e 65%), o desempenho dos métodos se altera. O método das k-médias continua sendo o mais eficiente, agora com 83,00% de acerto. As distâncias Euclidiana e Euclidiana Média foram agora as que tiveram maior destaque no agrupamento hierárquico junto ao algoritmo de Ward.

Considerando sete variáveis quantitativas (excluiu-se as variáveis com herdabilidade 70% ,65%, 60%), o algoritmo k-médias manteve alto desempenho (81,67%). A distância Euclidiana e de Gower associadas ao método de Ward foram as que tiveram maior representação.

Com seis variáveis quantitativas (excluiu-se as variáveis com herdabilidade 70% ,65%, 60% e 65%), a qualidade do agrupamento diminuiu, porém os destaques continuam sendo o método das k-médias e o algoritmo de Ward, que possui, por exemplo, 74,29% de acerto.

Trabalhando com as cinco menores taxas de herdabilidade (excluiu-se as variáveis com herdabilidade 70% ,65%, 60%, 55% e 50%), o método de Ligação Completa passa a ter desempenho superior que o de Ward em alguns casos, porém o algoritmo não hierárquico continua sendo o mais eficiente.

Com quatro variáveis quantitativas (excluiu-se as variáveis com herdabilidade 70% ,65%, 60%, 55% ,50% e 45%), o desempenho do método k-médias foi mais uma vez superior aos demais.

Ao utilizar três variáveis quantitativas (excluiu-se as variáveis com herdabilidade 70% ,65%, 60%, 55% ,50% ,45% e 40%), o k-médias e a distância de Gower combinada com o algoritmo de Ward foram os que se sobressaíram com 76,29% e 72,25% de acerto, respectivamente.

No último cenário da avaliação das distâncias e métodos de agrupamento utilizando somente variáveis quantitativas, observa-se que quando se trabalha com baixas herdabilidade (25% e 30%), a qualidade do agrupamento tende a ser menor do que quando tinha-se mais informação nas variáveis utilizadas. Dessa vez, a distância de Gower com o método de Ward conseguiu superar o método não hierárquico.

A seguir será apresentado o desempenho dos métodos e distâncias trabalhando no cenário de mistura de variáveis, ou seja, variáveis quantitativas e binárias.

3.4. Análise de nove variáveis quantitativas e uma variável binária

No capítulo anterior foi relatado que, como havia dez variáveis quantitativas, donde a herdabilidade varia de 25% até 70% entre as variáveis, seria categorizado em duas categorias sempre a última variável, ou seja, ao se trabalhar com nove variáveis quantitativas e uma binária, fica entendido que classificou-se a mais alta característica de herdabilidade (70%) de acordo com o percentil 50:

Após a categorização, pode-se empregar as distâncias combinadas (MINGOTI, 2005) apresentadas no capítulo anterior.

O método das k-médias não pode mais ser empregado já que trabalha estritamente com variáveis quantitativas.

A tabela 15 apresenta o desempenho dos algoritmos hierárquicos Ligação Completa e Ward e do não hierárquico k-protótipos para quatro valores distintos de γ.

Os protótipos iniciais adotados para utilizar o k-protótipos foram as duas primeiras observações do banco de dados.

As taxas de acerto iguais ou superiores a 80,00% estão em negrito com o propósito de destacar a eficiência do método e distância utilizados.

Tabela 15 – Taxa de desempenho utilizando nove variáveis quantitativas e uma binária

Métodos	% de acerto			
	$\gamma = 0,5$	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$
Euclidiana.Comb - Lig_Completa	79,25%	-	-	-
Euclidiana.Comb – Ward	78,50%	-	-	-
Euclidiana ao quadrado - Lig_Completa	74,75%	75,67%	-	74,13%
Euclidiana ao quadrado.Comb – Ward	73,63%	70,96%	73,83%	**81,08%**
Euclidiana Média.Comb - Lig_Completa	69,88%	-	-	-
Euclidiana Média.Comb – Ward	**80,79%**	-	-	-
Mahalanobis.Comb - Lig_Completa	-	-	-	-
Mahalanobis.Comb – Ward	-	-	-	-
Manhattan.Comb - Lig_Completa	-	71,42%	68,92%	-
Manhattan.Comb – Ward	**81,00%**	79,13%	-	-
Gower – Ligação Completa			-	
Gower – Ward			-	
k-protótipos - Euclidiana Combinada	77,00%	-	-	-
k-protótipos - Euclidiana ao quadrado Combinada	**82,17%**	**82,75%**	**83,21%**	79,08%
k-protótipos - Euclidiana Média Combinada	**80,88%**	-	-	-
k-protótipos - Mahalanobis Combinada	-	-	-	-
k-protótipos - Manhattan Combinada	**81,46%**	**80,17%**	-	-
k-protótipos – Gower			-	

Pela tabela acima observa-se que dependendo da distância utilizada e valor de *gamma*, tanto os métodos hierárquicos quanto o não hierárquico k-protótipos não são capazes de identificar as populações, inviabilizando a mensuração de acerto.

Essa confusão é decorrente do resultado agrupar os P_1 e P_2 no mesmo grupo, sendo que estes são os mais distintos do estudo.

Mesmo com algumas combinações sendo ineficazes, conclui-se que o k-protótipos foi mais eficiente que os métodos hierárquicos, principalmente ao ser utilizado junto a medida de dissimilaridade Euclidiana ao quadrado Combinada.

3.5. Análise de oito variáveis quantitativas e duas variáveis binárias

Trabalhando em um cenário com oito variáveis quantitativas e duas variáveis binárias (herdabilidade 70% e 65%), temos os seguintes resultados apresentados na tabela 9.

Tabela 16 – Taxa de desempenho utilizando oito variáveis quantitativas e duas binárias

Métodos	% de acerto			
	$\gamma = 0,5$	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$
Euclidiana.Comb - Lig_Completa	73,08%	68,04%	-	-
Euclidiana.Comb - Ward	75,29%	62,00%	61,92%	61,92%
Euclidiana ao quadrado - Lig_Completa	71,50%	70,92%	66,42%	72,25%
Euclidiana ao quadrado.Comb - Ward	77,33%	78,42%	77,54%	73,54%
Euclidiana Média.Comb - Lig_Completa	69,58%	-	-	-
Euclidiana Média.Comb - Ward	**80,96%**	62,04%	61,92%	61,92%
Mahalanobis.Comb - Lig_Completa	-	-	-	-
Mahalanobis.Comb - Ward	-	61,88%	61,92%	61,92%
Manhattan.Comb - Lig_Completa	58,67%	74,92%	70,58%	61,92%
Manhattan.Comb - Ward	**80,13%**	**81,71%**	62,00%	61,96%
Gower – Ligação Completa		61,92%		
Gower - Ward		61,92%		
k-protótipos - Euclidiana Combinada	77,79%	75,21%	75,21%	75,21%
k-protótipos - Euclidiana ao quadrado Combinada	**83,75%**	**82,75%**	**81,75%**	78,17%
k-protótipos - Euclidiana Média Combinada	79,25%	75,08%	75,08%	75,08%
k-protótipos - Mahalanobis Combinada	75,96%	76,08%	76,08%	76,08%
k-protótipos - Manhattan Combinada	**81,67%**	79,00%	73,38%	73,38%
k-protótipos - Gower		73,71%		

A tabela acima mostra o desempenho superior do k-protótipos frente aos métodos de Ward e Ligação Completa. Ao comparar somente os métodos hierárquicos, observa-se que o método de Ward foi mais eficiente que o Ligação Completa para a maioria das combinações realizadas.

3.6. Análise de sete variáveis quantitativas e três variáveis binárias

A seguir, a tabela mostra os resultados ao se usar sete variáveis quantitativas e três binárias (herdabilidade 70%,65% e 60%).

Tabela 17 – Taxa de desempenho utilizando sete variáveis quantitativas e três binárias

Métodos	% de acerto			
	$\gamma = 0,5$	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$
Euclidiana.Comb - Lig_Completa	-	67,63%	-	-
Euclidiana.Comb – Ward	**80,46%**	-	61,92%	61,92%
Euclidiana ao quadrado - Lig_Completa	-	73,08%	69,29%	-
Euclidiana ao quadrado.Comb – Ward	71,83%	73,25%	79,83%	79,42%
Euclidiana Média.Comb - Lig_Completa	63,67%	62,58%	-	-
Euclidiana Média.Comb – Ward	73,92%	-	61,92%	61,92%
Mahalanobis.Comb - Lig_Completa	-	-	-	61,92%
Mahalanobis.Comb – Ward	75,29%	66,00%	61,92%	61,92%
Manhattan.Comb - Lig_Completa	66,38%	74,71%	67,29%	-
Manhattan.Comb – Ward	75,13%	74,38%	62,08%	61,92%
Gower – Ligação Completa		66,08%		
Gower – Ward		61,92%		
k-protótipos - Euclidiana Combinada	78,04%	64,79%	63,21%	63,21%
k-protótipos - Euclidiana ao quadrado Combinada	**84,25%**	**83,67%**	**80,42%**	75,58%
k-protótipos - Euclidiana Média Combinada	78,42%	64,79%	63,21%	63,21%
k-protótipos - Mahalanobis Combinada	64,79%	64,79%	63,21%	63,21%
k-protótipos - Manhattan Combinada	78,46%	-	64,79%	64,79%
k-protótipos – Gower		63,21%		

O algoritmo k-protótipos associado a distância Euclidiana ao quadrado Combinada mais uma vez é destaque em relação aos demais. Ao utilizar γ=0,5, o percentual de acerto atinge 84,25%.

A distância Euclidiana Combinada junto ao método de Ward para γ=0,5 obteve o melhor desempenho dentre os hierárquicos (80,46%), porém, é ainda inferior ao k-protótipos.

A distância de Gower não obteve desempenho bom em nenhum dos casos utilizados.

3.7. Análise de seis variáveis quantitativas e quatro variáveis binárias

Ao se trabalhar em um cenário com seis variáveis quantitativas e quatro binárias (herdabilidade 70%, 65%, 60% e 55%), percebe-se que só o k-protótipos atingiu mais de 80% de acerto do agrupamento dos genótipos.

O método da Ligação Completa associado a distância Euclidiana ao quadrado para γ=0,5 e γ=3 foi um dos casos que não foi possível mensurar a taxa de acerto do agrupamento já que condensou no mesmo grupo os genitores.

Ao comparar somente os dois métodos hierárquicos, conclui-se que o método de Ward é mais eficiente que o Ligação Completa, porém não consegue atingir o mesmo patamar de acerto que o k-protótipos.

Tabela 18 – Taxa de desempenho utilizando seis variáveis quantitativas e quatro binárias

Métodos	% de acerto			
	γ = 0,5	γ = 1	γ = 2	γ = 3
Euclidiana.Comb - Lig_Completa	69,83%	65,04%	62,75%	62,75%
Euclidiana.Comb - Ward	76,29%	59,50%	59,21%	59,21%
Euclidiana ao quadrado - Lig_Completa	-	70,08%	70,92%	-
Euclidiana ao quadrado.Comb - Ward	71,67%	68,92%	75,21%	76,63%
Euclidiana Média.Comb - Lig_Completa	66,33%	64,29%	64,79%	64,79%
Euclidiana Média.Comb - Ward	72,38%	59,58%	59,21%	59,21%
Mahalanobis.Comb - Lig_Completa	-	-	62,75%	-
Mahalanobis.Comb - Ward	76,00%	65,79%	59,21%	59,21%
Manhattan.Comb - Lig_Completa	-	71,79%	72,00%	74,71%
Manhattan.Comb - Ward	71,50%	76,63%	66,17%	59,21%
Gower – Ligação Completa		74,67%		
Gower - Ward		59,21%		
k-protótipos - Euclidiana Combinada	79,83%	79,00%	79,00%	76,79%
k-protótipos - Euclidiana ao quadrado Combinada	77,91%	**80,95%**	**81,45%**	79,83%
k-protótipos - Euclidiana Média Combinada	**80,29%**	79,04%	79,04%	76,79%
k-protótipos - Mahalanobis Combinada	79,33%	79,46%	79,54%	77,71%
k-protótipos - Manhattan Combinada	**80,75%**	**80,04%**	78,33%	78,38%
k-protótipos - Gower		68,42%		

3.8. Análise de cinco variáveis quantitativas e cinco variáveis binárias

No cenário em que se trabalha com cinco variáveis quantitativas e cinco variáveis binárias o destaque também é o método k-protótipos associado a distância Euclidiana ao quadrado Combinada.

Para γ=0,5, o método de Ward é mais eficiente que o Ligação Completa, porém, alterando o parâmetro *gamma* para 1, o comportamento do primeiro passar a ser melhor.

O método k-protótipos possui resultados mais homogêneos que os algoritmos hierárquicos, em que o desempenho está acima de 70% para todas combinações estudadas

Tabela 19 – Taxa de desempenho utilizando cinco variáveis quantitativas e cinco binárias

Métodos	% de acerto			
	$\gamma = 0,5$	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$
Euclidiana.Comb - Lig_Completa	73,42%	71,92%	65,04%	-
Euclidiana.Comb - Ward	78,79%	63,13%	66,13%	66,08%
Euclidiana ao quadrado - Lig_Completa	67,79%	74,00%	-	73,71%
Euclidiana ao quadrado.Comb - Ward	73,83%	74,71%	78,58%	74,17%
Euclidiana Média.Comb - Lig_Completa	74,13%	65,42%	-	-
Euclidiana Média.Comb - Ward	**80,21%**	68,04%	66,13%	66,08%
Mahalanobis.Comb - Lig_Completa	61,38%	71,25%	75,00%	73,00%
Mahalanobis.Comb - Ward	72,67%	71,63%	60,46%	66,08%
Manhattan.Comb - Lig_Completa	70,00%	69,79%	68,33%	72,96%
Manhattan.Comb - Ward	74,21%	76,46%	68,67%	66,08%
Gower – Ligação Completa	61,54%			
Gower - Ward	66,08%			
k-protótipos - Euclidiana Combinada	**80,38%**	74,13%	74,13%	72,46%
k-protótipos - Euclidiana ao quadrado Combinada	**80,13%**	**81,58%**	**81,08%**	77,45%
k-protótipos - Euclidiana Média Combinada	**80,54%**	74,13%	74,13%	72,46%
k-protótipos - Mahalanobis Combinada	78,58%	74,13%	74,13%	72,46%
k-protótipos - Manhattan Combinada	**81,38%**	**80,04%**	74,13%	74,13%
k-protótipos - Gower	72,67%			

3.9. Análise de quatro variáveis quantitativas e seis variáveis binárias

A tabela a seguir apresenta o desempenho dos algoritmos de agrupamento e medidas de distância combinadas para o cenário em que existe quatro variáveis quantitativas e seis variáveis binárias (herdabilidade 70%, 65%, 60%, 55%, 50% e 45%).

O algoritmo k-protótipos se destaca mais uma vez com os maiores níveis de eficiência e mantendo seus resultados mais homogêneos que os métodos hierárquicos, onde há uma brusca mudança ao se variar a medida de distância ou o método utilizado ou inclusive o próprio *gamma*.

Algumas das combinações de distância e método hierárquico mais uma vez não foram capazes de identifcar os grupos, como por exemplo ocorreu com o método

de Ligação Completa utilizando as distâncias Euclidiana ao quadrado Combinada, Euclidiana Média Combinada e Manhattan Combinada.

Tabela 20 – Taxa de desempenho utilizando quatro variáveis quantitativas e seis binárias

Métodos	% de acerto			
	$\gamma = 0,5$	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$
Euclidiana.Comb - Lig_Completa	76,50%	59,67%	68,67%	-
Euclidiana.Comb - Ward	75,63%	73,96%	67,54%	57.88%
Euclidiana ao quadrado - Lig_Completa	68,17%	-	-	-
Euclidiana ao quadrado.Comb - Ward	73,67%	75,96%	71,13%	79,83%
Euclidiana Média.Comb - Lig_Completa	-	69,88%	-	68,41%
Euclidiana Média.Comb - Ward	75,79%	69,00%	67,54%	57,88%
Mahalanobis.Comb - Lig_Completa	73,29%	66,88%	-	72,92%
Mahalanobis.Comb - Ward	74,13%	-	71,83%	57,88%
Manhattan.Comb - Lig_Completa	-	72,88%	62,37%	-
Manhattan.Comb - Ward	76,67%	76,04%	68,83%	70,92%
Gower – Ligação Completa	-			
Gower - Ward	-			
k-protótipos - Euclidiana Combinada	79,29%	78,46%	78,46%	78,46%
k-protótipos - Euclidiana ao quadrado Combinada	**81,21%**	**82,63%**	**80,54%**	78,83%
k-protótipos - Euclidiana Média Combinada	79,46%	78,46%	78,46%	78,46%
k-protótipos - Mahalanobis Combinada	78,50%	78,33%	78,33%	78,33%
k-protótipos - Manhattan Combinada	**81,54%**	77,92%	77,92%	77,92%
k-protótipos - Gower	77,96%			

3.10. Análise de três variáveis quantitativas e sete variáveis binárias

A seguir encontra-se a tabela 21, referente ao desempenho dos métodos e distâncias avaliados para três variáveis quantitativas e sete variáveis binárias (herdabilidade 70%, 65%, 60%, 55%, 50%, 45% e 40%).

Dentre os métodos hierárquicos, o destaque vai para a distância de Mahalanobis Combinada associada ao método de Ward e γ=0,5 com 80,96% de aproveitamento.

A maior performance ainda é do k-protótipos associado a distância Euclidiana ao quadrado Combinada. Mesmo com apenas 3 variáveis quantitativas e outras 7 binárias, o algoritmo mostra-se bastante eficaz.

Tabela 21 – Taxa de desempenho utilizando três variáveis quantitativas e sete binárias

Métodos	% de acerto			
	$\gamma = 0,5$	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$
Euclidiana.Comb - Lig_Completa	-	69,96%	66,08%	56,71%
Euclidiana.Comb – Ward	74,67%	59,17%	59,21%	59,21%
Euclidiana ao quadrado - Lig_Completa	69,79%	-	-	-
Euclidiana ao quadrado.Comb – Ward	74,58%	72,38%	74,50%	73,21%
Euclidiana Média.Comb - Lig_Completa	-	77,54%	66,08%	66,08%
Euclidiana Média.Comb – Ward	74,83%	61,92%	59,21%	59,21%
Mahalanobis.Comb - Lig_Completa	-	-	66,00%	63,21%
Mahalanobis.Comb – Ward	**80,96%**	61,58%	59,21%	59,21%
Manhattan.Comb - Lig_Completa	74,08%	-	-	57,67%
Manhattan.Comb – Ward	76,83%	71,33%	58,04%	59,21%
Gower – Ligação Completa		60,42%		
Gower – Ward		68,54%		
k-protótipos - Euclidiana Combinada	**80,92%**	73,92%	73,92%	73,92%
k-protótipos - Euclidiana ao quadrado Combinada	**80,88%**	**81,88%**	79,33%	76,17%
k-protótipos - Euclidiana Média Combinada	**81,00%**	73,92%	73,92%	73,92%
k-protótipos - Mahalanobis Combinada	79,17%	73,92%	73,92%	73,92%
k-protótipos - Manhattan Combinada	**81,54%**	79,63%	73,92%	73,92%
k-protótipos – Gower		73,92%		

3.11. Análise de duas variáveis quantitativas e oito variáveis binárias

No cenário mais extremo desse estudo, avaliou-se o caso de existir apenas duas variáveis quantitativas e oito variáveis binárias (herdabilidade 70%, 65%, 60%, 55%, 50%, 45%, 40% e 35%).

Tabela 22 – Taxa de desempenho utilizando duas variáveis quantitativas e oito binárias

Métodos	% de acerto			
	$\gamma = 0,5$	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$
Euclidiana.Comb - Lig_Completa	62,83%	-	63,17%	-
Euclidiana.Comb - Ward	74,83%	-	73,67%	71,67%
Euclidiana ao quadrado - Lig_Completa	-	65,46%	63,00%	-
Euclidiana ao quadrado.Comb - Ward	64,21%	56,50%	71,13%	72,41%
Euclidiana Média.Comb - Lig_Completa	68,75%	67,08%	-	-
Euclidiana Média.Comb - Ward	69,17%	-	67,67%	71,67%
Mahalanobis.Comb - Lig_Completa	68,04%	72,17%	-	64,79%
Mahalanobis.Comb - Ward	75,21%	74,83%	71,38%	67,25%
Manhattan.Comb - Lig_Completa	67,21%	72,88%	-	-
Manhattan.Comb - Ward	-	69,25%	73,50%	67,92%
Gower – Ligação Completa	-			
Gower - Ward	-			
k-protótipos - Euclidiana Combinada	79,58%	79,63%	79,58%	79,58%
k-protótipos - Euclidiana ao quadrado Combinada	77,63%	**80,29%**	79,75%	79,63%
k-protótipos - Euclidiana Média Combinada	79,88%	79,58%	79,88%	79,58%
k-protótipos - Mahalanobis Combinada	79,88%	79,88%	79,88%	79,67%
k-protótipos - Manhattan Combinada	**80,50%**	79,79%	79,79%	79,83%
k-protótipos - Gower	75,13%			

Nesse caso, somente o k-protótipos obteve resultados superiores a 80%. O melhor desempenho entre os métodos hierárquicos foi quando utilizou-se a distância de Mahalanobis Combinada no algoritmo de Ward para *gamma* 0,5, em que a taxa de acerto foi de 75,21%.

4. CONCLUSÕES

Como foi exposto e analisado, o desempenho dos métodos de agrupamento não hierárquico k-médias e k-protótipos foram superiores aos algoritmos hierárquicos testados.

No primeiro cenário, onde analisou-se somente variáveis quantitativas, o k-médias só não foi o melhor para o caso extremo em que trabalhou-se apenas duas variáveis.

Ao estudar mistura de variáveis, o algoritmo k-protótipos, que também é da classe dos não hierárquicos, conseguiu melhor desempenho, principalmente sendo utilizado junto a distância Euclidiana ao quadrado Combinada e *gamma* igual a 0,5 ou 1.

Conclui-se também que é interessante incluir variáveis binárias em estudos de diversidade genética, pois quando utilizou-se destas, as taxas de desempenho foram superiores aos cenários que utilizaram somente informações das variáveis contínuas.

Métodos hierárquicos não se adequaram bem a esse estudo, o que permite indicar para trabalhos futuros no campo da Divesidade Genética, o uso de métodos não hierárquicos, sendo que se houver mistura de variáveis, que utilize-se o k-protótipos.

Esses métodos citados possuem comportamento bastante eficazes ao serem utilizados em bases de dados grande, como é o caso em estudo, que possui 2400 obervações e dez variáveis.

5. REFERÊNCIAS

BARBOSA, C. D; VIANA, A.P; QUINTAL, S.S.R; PEREIRA, M.G. Artificial neural network analysis of genetic diversity in Carica papaya L.. **Crop Breeding and Applied Biotechnology** (Impresso), v. 11, p. 224-231, 2011.

BARROSO, N. C. **Categorização de dados quantitativos para estudos de diversidade genética**. Viçosa, 2010. 99p. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa.

BHERING, L. L; LAVIOLA, B. G ; ROSADO, T. B ; Alves, A. A. **Metodologias de avaliação conjunta da diversidade genética baseada em informações Agronômicas e moleculares aplicadas a pinhão manso.** 2011, Buzios. SBMP, 2011. Referências adicionais: Classificação do evento: Brasil/ Português.

BOLDT, A. S. **Diversidade genética , adaptabilidade e estabilidade de genótipos de soja no Mato Grosso**. Viçosa, 2011. 205p. Dissertação (Mestrado em Genética e Melhoramento) – Universidade Federal de Viçosa, 2011

BRITO, G ; ASSAF ,N. A; CORRAR, L. J. **Sistema de Classificação de Risco de Crédito: uma aplicação a companhias abertas no Brasil**. Revista Contabilidade & Finanças (Impresso), v. 20, p. 28-43, 2009.

CARDELINO, R.; OSÓRIO, J. C. S. **Melhoramento Animal para Agronomia, Veterinária e Zootecnia**. Pelotas: Editora UFPel., 1999. 153p.

CARVALHO, L. P; LANZA, M. A.; FALIERI, J.; SANTOS, J. W. **Análise da diversidade genética entre acessos do banco ativo de germoplasma de algodão**. Pesquisa Agropecuária Brasileira, v.38, n.10, p.1149-1155, 2003.

COELHO, C. M. M; COIMBRA, J. L. M; SOUZA, C. A; BOGO, A; GUIDOLIN, A. F. **Divergência Genética em acessos de feijão (Phasolus vulgaris l.)**. Ciência Rural, v. 37, p. 1241-1247, 2007.

CRUZ, C.D. **Programa Genes**: Biometria. Editora UFV. Viçosa (MG). 382p. 2006.

CRUZ, C.D; FERREIRA, F.M; PESSONI, L.A. **Biometria Aplicada ao estudo da diversidade genética**. Viçosa, 2008. 539p.

CRUZ, C. D. **Programa genes (versão Windows): aplicativo computacional em genética e estatística.** Viçosa: UFV, 2008.

CRUZ, C. D.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético**. v. 2, 2 ed., Viçosa: UFV, 623p. 2003.

CRUZ, C. D.; FERREIRA, F. M.; PESSONI, L. A. **Biometria aplicada ao estudo da diversidade genética**. Visconde do Rio Branco: Suprema, 620p. 2011.

CRUZ, C. D.; REGAZZI, A. J.; CARNEIRO, P. C. S. **Modelos Biométricos Aplicados ao Melhoramento Genético**. v.1, 3 ed., Viçosa: UFV, 480p. 2004.

ELIAS ; GONCALVES; VIDIGAL, M. C. **Variabilidade genética em germoplasma tradicional de feijão-preto em Santa Catarina**. Pesquisa Agropecuária Brasileira, v. 42, p. 1443-1449, 2007.

FARIA, P. N. **Avaliação de métodos para determinação do número ótimo de *clusters* em estudo de divergência genética entre acessos de pimenta**. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa. Orientador: Paulo Roberto Cecon, Viçosa, 67f. 2009.

FARIA, P. N; CECON, P. R; SILVA, A. R ; FINGER, F. L ; SILVA, F. F ; CRUZ, C. D ; SAVIO, F. L . **Métodos de agrupamento em estudo de divergência genética de pimentas**. Horticultura Brasileira (Impresso), v. 30, p. 428-432, 2012.

FRANCO, J.; CROSSA, J.; DÍAZ, J.; TABA, S.; VILLASEÑOR, J; EBERHART, A. **A sequential clustering for classifying gene bank accessions**. Crop Science, 37: 1656-1662, 1997.

FERREIRA, F.M**. Diversidade em populações simuladas com base em locos multialélicos**. Viçosa, 2007. 177p. Tese (Doutorado em Genética e Melhoramento) – Universidade Federal de Viçosa, 2007.

GOWER, J. C. **Some distance properties of latent root and vector methods used in multivariate analysis**. *Biometrika 63,* 315-28, 1966.

GOWER, J. C. **A General Coefficient of Similarity and Some of its Properties**. BioMetrics, 27, pp. 857-874, 1971.

GUHA, S; TASTOGI, R; SHIM, K. **Rock: a Robust Clustering Algorithm for Categorical Atributtes**. In: Proceedings of the 15th International Conference on Data Engineering, pp. 512-521, Washington, USA, 1999.

HUANG, Z. **Clustering Large Data Sets with Mixed Numeric and Categorical Values**. In Proceedings of The First Pacific-Asia Conference on Knowledge Discovery and Data Mining, Singapore, World Scientific, 1997.

HUANG, Z. **A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining**. Research Issues on Data Mining and Knowledge Discovery, 1997.

HUANG, Z. **Extensions to the k-means algorithm for clustering large data sets with categorical values**. — Data Mining Knowl. Discov., Vol. 2, No. 2, pp. 283–304, 1998.

MAHALANOBIS, P. C. **On the generalized distance in statistics.** Proceedings of The National Institute of Sciences of India, v.12, p.49-55, 1936.

MACQUEEN, J. B. (1967). **Some methods for classification and analysis of multivariate observations**. In Cam, L. M. L. and Neyman, J., editors, Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 281–297.University of California Press.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada**: uma abordagem aplicada. Belo Horizonte: Editora UFMG, 297p. 2005.

OLIVEIRA, M.M.; ROTA, E. L.; DIONELLO, N. J. L.; AITA, M. F. **Herdabilidade e correlações genéticas do perímetro escrotal com características produtivas em bovinos de corte**: Revisão. Revista Brasileira de Agrociência, v. 13, p. 141-146, 2007.

PIMENTEL, E. P; FRANÇA, V. F; OMAR, N. **A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização**. In: XIV SBIE - Simpósio Brasileiro de Informática na Educação, 2003, Rio de Janeiro / RJ. Anais do XIV Simpósio Brasileiro de Informática na Educação, 2003. p. 523-532.

R DEVELOPMENT CORE TEAM.**R: A language and environment forstatistical computing**. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. 2012

SILVA, A. R., **Métodos de agrupamento: avaliação e aplicação ao Estudo de divergência genética em acessos de alho**. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa. Orientador: Paulo Roberto Cecon, Viçosa, 83f. 2012.

SILVEIRA, T. C; LEDO, C. A. S; TAVARES FILHO, L. F. Q; ALVES, A. A. C; SANTOS, A. S. **Diversidade genética entre cultivares de mandioca e espécies silvestres de manihot mediante caracterização morfológica**. In: 3ª Jornada Científica, 2009, Cruz das Almas. Anais, 2009.

SIMÕES, R. F. **Localização Industrial e Relações Intersetoriais: uma análise de fuzzy cluster para Minas Gerais**. Campinas, 2003. 183p. Tese (Doutorado em Economia) - Universidade de Campinas.

VIANA, A. P ; PEREIRA, T. N. S ; PEREIRA, M. G; SOUZA, M. M; MALDONADO, J.F.M; AMARAL, J. A. T. **Diversidade Genética entre Genótipos**

Comerciais de Maracujazeiro-Amarelo (Passiflora edulis f. flavicarpa) e entre Espécies de Passifloras Nativas Determinadas por Marcadores RAPD. Revista Brasileira de Fruticultura, Jaboticabal, v. 25, n.3, p. 489-493, 2003.

APÊNDICE

A – Script das análises de agrupamento hierárquicas no *Software* R

```
agrup.h<-function(dados,nquant,ngrupos)
{
library(StatMatch)
k=ngrupos
dados=as.matrix(dados)
pop=read.table("pop12.txt")
distancia=matrix(0,nrow(dados),nrow(dados))
colnames(distancia)=c(seq(1,nrow(dados)))
rownames(distancia)=c(seq(1,nrow(dados)))
diag_matrix_inv_cov=diag(solve(cov(dados[,1:nquant])))
diag_matrix_inv_cov=diag(diag_matrix_inv_cov)
matrix_inv_cov=solve(cov(dados[,1:nquant]))

readline("As opções de distância são:1.Euclidiana
;2.Euclidiana ao quadrado;3.Euclidiana
Média;4.Mahalanobis;5.Manhattan;6.Gower;7.Euclidiana.Comb
;8.Euclidiana ao quadrado.Comb;9.Euclidiana
Média.Comb;10.Mahalanobis.Comb;11.Manhattan.Comb")
opcao<-readline("Informe a distância desejada:  ")

if(opcao=="Euclidiana")
{
      distancia=dist(dados[,1:nquant])

#print(as.dist(distancia))

}

if(opcao=="Euclidiana.Comb")
{
```

```
gama<-as.numeric(readline("Informe o valor de gama:  "))

for (i in 1:nrow(dados))

      {

     for (j in 1:nrow(dados))

       {
          distancia[i,j]=sqrt(sum((dados[i,1:nquant]-
dados[j,1:nquant])^2))+gama*sum(dados[i,(nquant+1):ncol(d
ados)]!=dados[j,(nquant+1):ncol(dados)])

      }
          }

distancia=as.dist(distancia)

#print(distancia)
}

if(opcao=="Euclidiana ao quadrado")
{
     distancia=(dist(dados[,1:nquant]))^2

     distancia=as.dist(distancia)

#print(distancia)

}

if(opcao=="Euclidiana ao quadrado.Comb")
{
gama<-as.numeric(readline("Informe o valor de gama:  "))
```

```
for (i in 1:nrow(dados))

      {

     for (j in 1:nrow(dados))

       {
          distancia[i,j]=(sum((dados[i,1:nquant]-
dados[j,1:nquant])^2))+gama*sum(dados[i,(nquant+1):ncol(d
ados)]!=dados[j,(nquant+1):ncol(dados)])

      }
          }
distancia=as.dist(distancia)

#print(distancia)
}

if(opcao=="Euclidiana Média")
{
for (i in 1:nrow(dados))

      {

     for (j in 1:nrow(dados))

       {
          distancia[i,j]=sqrt(((t(dados[i,1:nquant])-
t(dados[j,1:nquant]))%*%diag_matrix_inv_cov%*%((dados[i,1
:nquant])-(dados[j,1:nquant]))))

     }
```

```
         }
distancia=as.dist(distancia)

#print(distancia)
}
if(opcao=="Euclidiana Média.Comb")
{
gama<-as.numeric(readline("Informe o valor de gama:  "))
for (i in 1:nrow(dados))
      {

     for (j in 1:nrow(dados))

       {
          distancia[i,j]=sqrt(((t(dados[i,1:nquant])-
t(dados[j,1:nquant]))%*%diag_matrix_inv_cov%*%((dados[i,1
:nquant])-
(dados[j,1:nquant]))))+gama*sum(dados[i,(nquant+1):ncol(d
ados)]!=dados[j,(nquant+1):ncol(dados)])
          }

          }

distancia=as.dist(distancia)

#print(distancia)
}

if(opcao=="Mahalanobis")
{
     distancia=sqrt(mahalanobis.dist(dados[,1:nquant]))

     distancia=as.dist(distancia)
```

```
#print(distancia)
}

if(opcao=="Mahalanobis.Comb")
{
gama<-as.numeric(readline("Informe o valor de gama:  "))

for (i in 1:nrow(dados))

      {

     for (j in 1:nrow(dados))

       {
          distancia[i,j]=sqrt(((t(dados[i,1:nquant])-
t(dados[j,1:nquant]))%*%matrix_inv_cov%*%((dados[i,1:nqua
nt])-
(dados[j,1:nquant]))))+gama*sum(dados[i,(nquant+1):ncol(d
ados)]!=dados[j,(nquant+1):ncol(dados)])

      }

          }
distancia=as.dist(distancia)

#print(distancia)

}

if(opcao=="Manhattan")
{

     distancia=dist(dados[,1:nquant],method="manhattan")
     distancia=as.dist(distancia)
```

```
#print(distancia)

}

if(opcao=="Manhattan.Comb")
{

gama<-as.numeric(readline("Informe o valor de gama:  "))
lambda<-as.numeric(readline("Informe o valor de lambda:
"))
     for (i in 1:nrow(dados))

      {

     for (j in 1:nrow(dados))

       {
          distancia[i,j]=(sum(abs(dados[i,1:nquant]-
dados[j,1:nquant])^lambda))^(1/lambda)+gama*sum(dados[i,(
nquant+1):ncol(dados)]!=dados[j,(nquant+1):ncol(dados)])
      }

          }

distancia=as.dist(distancia)

#print(distancia)
}

if(opcao=="Gower")
{
      distancia=as.dist(gower.dist(dados))
```

```
#print(as.dist(distancia))
}

readline("As opções de agrupamento são:1.Ligação Simples;2.Ligação Completa;3.UPGMA;4.Ward")
opcao<-readline("Informe o método de agrupamento desejado:  ")

if(opcao=="Ligação Simples")
{

agrup<-hclust(distancia,method ="single")
plot(agrup,main="Dendograma - Método Ligação Simples",ylab="Distância",xlab="Observações",sub="")
rect.hclust(agrup, k = k)
id<-seq(1,nrow(dados))
grupos<-cutree(agrup, k=k)
grupos1<-cbind(id,pop,grupos)
grupos1

#print(grupos1)

write.table(grupos1,file='grupos1.xls',row.names=F)

d.cof <- cophenetic(agrup)

print("Coeficiente de Correlação Cofenética")
print(cor(distancia,d.cof))


}

if(opcao=="Ligação Completa")
{
```

```
agrup<-hclust(distancia,method ="complete")
plot(agrup,main="Dendograma - Método Ligação
Completa",ylab="Distância",xlab="Observações",sub="")
rect.hclust(agrup, k = k)
id<-seq(1,nrow(dados))
grupos<-cutree(agrup, k=k)
grupos1<-cbind(id,pop,grupos)
grupos1

#print(grupos1)

write.table(grupos1,file='grupos1.xls',row.names=F)


d.cof <- cophenetic(agrup)

print("Coeficiente de Correlação Cofenética")
print(cor(distancia,d.cof))


}

if(opcao=="UPGMA")
{

agrup<-hclust(distancia,method ="average")
plot(agrup,main="Dendograma - Método
UPGMA",ylab="Distância",xlab="Observações",sub="")
rect.hclust(agrup, k = k)
id<-seq(1,nrow(dados))
grupos<-cutree(agrup, k=k)
grupos1<-cbind(id,pop,grupos)
grupos1
```

```
#print(grupos1)

write.table(grupos1,file='grupos1.xls',row.names=F)

d.cof <- cophenetic(agrup)

print("Coeficiente de Correlação Cofenética")
print(cor(distancia,d.cof))

}

if(opcao=="Ward")
{

agrup<-hclust(distancia,method ="ward")
plot(agrup,main="Dendograma - Método
Ward",ylab="Distância",xlab="Observações",sub="")
rect.hclust(agrup, k = k)
id<-seq(1,nrow(dados))
grupos<-cutree(agrup, k=k)
grupos1<-cbind(id,pop,grupos)
grupos1

#print(grupos1)

write.table(grupos1,file='grupos1.xls',row.names=F)

d.cof <- cophenetic(agrup)

print("Coeficiente de Correlação Cofenética")
print(cor(distancia,d.cof))
}
}
```

B – Script do algoritmo K-protótipos (distância proposta pelo autor HUANG(1997)) implementado em R

```
x<- #base de dados
nquant<- #número de variáveis quantitativas na base de dados
nquali<- #número de variáveis qualitativas na base de dados
nprot<-  #número de protótipos
gama<-   #valor adotado para o peso das variáveis qualitativas

prot<-matrix(c(),nrow=nprot,ncol=ncol(x),byrow=T)       #chutes
iniciais

indprod<-seq(1,nprot)
prot1<-cbind(indprod,prot)

print(prot1)

clustership<-matrix(NA,nrow=nrow(x),ncol=1)
grupo_teste<-matrix(0,50,ncol=nrow(x))
nmud<-matrix(0,50,ncol=1)
nmud[1,1]<-nrow(x)
h<-2

system.time(
repeat
{
  for(i in 1:nrow(x))
  {
  dist_minima<-sum((x[i,1:nquant]-
prot1[1,2:(nquant+1)])^2)+gama*sum(x[i,(nquant+1):ncol(x)]!=pr
ot1[1,(nquant+2):ncol(prot1)])
  grupo<-1
   for(j in 2:nprot)
   {
```

```
    dist<-sum((x[i,1:nquant]-
prot1[j,2:(nquant+1)])^2)+gama*sum(x[i,(nquant+1):ncol(x)]!=pr
ot1[j,(nquant+2):ncol(prot1)])
     if(dist<dist_minima)
     {
      dist_minima<-dist
      grupo<-j
     }
     else
     grupo<-grupo
    }
   clustership[i]<-grupo
   }
  clustership
  g<-intersect(clustership, clustership)
  g_ord<-g[order(g)]
  dados_ord<-cbind(clustership,x)
  dados_ord<-dados_ord[order(dados_ord[,1]),]
  prot_at_quant<-by(dados_ord[,2:(ncol(x)+1)],dados_ord[,1],
function(x) mean(x[,1:nquant]))
  prot_at_quant=as.list(prot_at_quant)
  prot_at_quant<-do.call("rbind",prot_at_quant)
  prot_at_quali<-by(dados_ord[,2:(ncol(x)+1)],dados_ord[,1],
function(x) round(mean(x[,(nquant+1):ncol(x)])))
  prot_at_quali=as.list(prot_at_quali)
  prot_at_quali<-do.call("rbind",prot_at_quali)
  prot_at<-cbind(prot_at_quant,prot_at_quali)

  #Atualização dos prototipos
  s=1
  for(i in  g_ord)
  {
  prot1[i,2:(ncol(x)+1)]<-prot_at[s,]
  s<-s+1
  }
  print(prot1)
  #
```

```
  grupo_teste[h,]<-t(clustership)
  nmud[h]<-nrow(x)-sum(grupo_teste[h,]==grupo_teste[h-1,])
  if(sum(grupo_teste[h,]-grupo_teste[h-1,])==0)
  break
  h<-h+1
  mudanca=nmud[1:length(nmud)-1]-nmud[2:length(nmud)]
  }

)


plot(seq(1,length(mudanca),1),nmud[2:length(nmud)],type="l",xl
ab="Iterações",ylab="Número de mudanças")


#################################################################
#########################################################

####################
## Função Verifica ##
####################

x_pop=read.table("pop_identidade.txt")
x_pop=x_pop[-c(401:600),1]

verifica=cbind(x_pop,clustership)
a=table(verifica[,1],verifica[,2])

 g1=sum(a[c(1,3,4,5,6,7),1])

 g2=sum(a[c(2,8,9,10,11,12),2])

 #g3=sum(a[c(3),3])
 acerto_total=(g1+g2)/2400
 acerto_total
 print(h-1)
```