UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Vinicius Campista Brum

Análise de Agrupamento e Estabilidade para Aquisição e Validação de Conhecimento em Bases de Dados de Alta Dimensionalidade

Juiz de Fora

2015

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Vinicius Campista Brum

Análise de Agrupamento e Estabilidade para Aquisição e Validação de Conhecimento em Bases de Dados de Alta Dimensionalidade

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Orientador: Itamar Leite de Oliveira
Coorientador: Wagner Antonio Arbex

Juiz de Fora
2015

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Brum, Vinicius Campista.
Análise de Agrupamento e Estabilidade para Aquisição e Validação de Conhecimento em Bases de Dados de Alta Dimensionalidade / Vinicius Campista Brum. -- 2015.
110 f. : il.

Orientador: Itamar Leite de Oliveira
Coorientador: Wagner Antonio Arbex
Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós-Graduação em Ciência da Computação, 2015.

1. Análise de agrupamento. 2. Análise de estabilidade. 3. Algoritmo genético. 4. GWAS. I. Oliveira, Itamar Leite de, orient. II. Arbex, Wagner Antonio, coorient. III. Título.

Vinicius Campista Brum

Análise de Agrupamento e Estabilidade para Aquisição e Validação de Conhecimento em Bases de Dados de Alta Dimensionalidade

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Aprovada em 28 de Agosto de 2015.

BANCA EXAMINADORA

Prof. Dr. Itamar Leite de Oliveira - Orientador
Universidade Federal de Juiz de Fora

Prof. Dr. Wagner Antonio Arbex
Universidade Federal de Juiz de Fora

Prof. Dr. Carlos Cristiano Hasenclever Borges
Universidade Federal de Juiz de Fora

Prof. Dr. Marcelo Costa Pinto e Santos
IF Sudeste MG – Campus Juiz de Fora

Dedico este trabalho aos meus pais, Edmundo e Ângela Brum, ao meu irmão, Eryck, e aos meus avós, José e Iraci Brum e Geraldina Campista.

AGRADECIMENTOS

Agradeço à minha família, especialmente aos meus pais, Edmundo e Ângela, e ao meu irmão, Eryck. Sem o incentivo, o apoio e a compreensão deles, este trabalho não teria sido possível.

Aos professores do Programa de Pós-Graduação em Ciência da Computação, que contribuíram para minha formação acadêmica. Especialmente aos meus orientadores, Prof. Itamar Oliveira e Prof. Wagner Arbex, pela paciência, dedicação e colaboração durante a realização deste trabalho.

Aos amigos Wadson Martins, Camillo Lellis, Jacimar Tavares, Vitor Freitas, Gustavo Henrique, Rodrigo Martins, Camila Campos e Alessandra Gomes, pela convivência, colaboração e parceria durante o mestrado.

Ao Centro Nacional de Pesquisa de Gado de Leite (Embrapa Gado de Leite) da Empresa Brasileira de Pesquisa Agropecuária e ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Juiz de Fora pelo fornecimento de infraestrutura necessária para condução deste trabalho. Especialmente ao Prof. Marcos Vinicius da Silva, por ceder o conjunto de dados utilizado neste trabalho, à Katia Cristina dos Santos e aos demais colegas do Laboratório de Bioinformática e Genômica Animal, pela colaboração.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro durante a realização deste trabalho.

"A pessimist sees the difficulty in every opportunity; an optimist sees the opportunity in every difficulty."

Winston Churchill

RESUMO

Análise de agrupamento é uma tarefa descritiva e não-supervisionada de mineração de dados que utiliza amostras não-rotuladas com o objetivo de encontrar grupos naturais, isto é, grupos de amostras fortemente relacionadas de forma que as amostras que pertençam a um mesmo grupo sejam mais similares entre si do que amostras em qualquer outro grupo. Avaliação ou validação é considerada uma tarefa essencial dentro da análise de agrupamento. Essa tarefa apresenta técnicas que podem ser divididas em dois tipos: técnicas não-supervisionadas ou de validação interna e técnicas supervisionadas ou de validação externa. Trabalhos recentes introduziram uma abordagem de validação interna que busca avaliar e melhorar a estabilidade do algoritmo de agrupamento por meio de identificação e remoção de amostras que são consideradas prejudiciais e, portanto, deveriam ser estudadas isoladamente. Por meio de experimentos foi identificado que essa abordagem apresenta características indesejáveis que podem resultar em remoção de todo um grupo e ainda não garante melhoria de estabilidade. Considerando essas questões, neste trabalho foi desenvolvida uma abordagem mais ampla utilizando algoritmo genético para análise de agrupamento e estabilidade de dados. Essa abordagem busca garantir melhoria de estabilidade, reduzir o número de amostras para remoção e permitir que o usuário controle o processo de análise de estabilidade, o que resulta em maior aplicabilidade e confiabilidade para tal processo. A abordagem proposta foi avaliada utilizando diferentes algoritmos de agrupamento e diferentes bases de dados, sendo que uma base de dados genotípicos também foi utilizada com o intuito de aquisição e validação de conhecimento. Os resultados mostram que a abordagem proposta é capaz de garantir melhoria de estabilidade e também é capaz de reduzir o número de amostras para remoção. Os resultados também sugerem a utilização da abordagem como uma ferramenta promissora para aquisição e validação de conhecimento em estudos de associação ampla do genoma (GWAS). Este trabalho apresenta uma abordagem que contribui para aquisição e validação de conhecimento por meio de análise de agrupamento e estabilidade de dados.

Palavras-chave: Análise de agrupamento. Análise de estabilidade. Algoritmo genético. GWAS.

ABSTRACT

Clustering analysis is a descriptive and unsupervised data mining task, which uses non-labeled samples in order to find natural groups, i.e. groups of closely related samples such that samples within the same cluster are more similar than samples within the other clusters. Evaluation and validation are considered essential tasks within the clustering analysis. These tasks present techniques that can be divided into two kinds: unsupervised or internal validation techniques and supervised or external validation techniques. Recent works introduced an internal clustering validation approach to evaluate and improve the clustering algorithm stability through identifying and removing samples that are considered harmful and therefore they should be studied separately. Through experimentation, it was identified that this approach has two undesirable characteristics, it can remove an entire cluster from dataset and still decrease clustering stability. Taking into account these issues, in this work a broader approach was developed using genetic algorithm for clustering and data stability analysis. This approach aims to increase stability, to reduce the number of samples for removal and to allow the user control the stability analysis process, which gives greater applicability and reliability for such process. This approach was evaluated using different kinds of clustering algorithm and datasets. A genotype dataset was also used in order to knowledge acquisition and validation. The results show the approach proposed in this work is able to increase stability, and it is also able to reduce the number of samples for removal. The results also suggest the use of this approach as a promising tool for knowledge acquisition and validation on genome-wide association studies (GWAS). This work presents an approach that contributes for knowledge acquisition and validation through clustering and data stability analysis.

Keywords: Clustering analysis. Data stability analysis. Genetic algorithm. GWAS.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES

AG	Algoritmo Genético.
CH	Índice de Calinski-Harabasz.
CSV	*Cluster Stability Variance.*
CVNN	*Clustering Validation index based on Nearest Neighbors.*
DBSCAN	*Density-Based Spatial Clustering of Applications with Noises.*
DI	Índice de Dunn.
EM	*Expectation-Maximization.*
GWAS	*Genome-Wide Association Study.*
HDDC	*High-Dimensional Data Clustering.*
SOM	*Self-Organizing Map.*
SNP	*Single Nucleotide Polymorphisms.*

SUMÁRIO

1 INTRODUÇÃO

Análise de agrupamento é uma tarefa descritiva e não-supervisionada de mineração de dados que utiliza amostras não-rotuladas com o objetivo de encontrar grupos naturais, isto é, grupos de amostras fortemente relacionadas de forma que as amostras que pertençam a um mesmo grupo sejam mais similares entre si do que amostras em qualquer outro grupo (DUDA et al., 2001; TAN et al., 2006; LIU et al., 2013).

Existem diversas razões para a utilização de tarefas de aprendizado não-supervisionado, dentre elas podemos destacar algumas mais importantes. Primeiro, a necessidade de rotular grandes bases de dados, o que poderia ser muito custoso sem o auxílio dessas tarefas. Segundo, sua utilização como uma tarefa preliminar para outras técnicas de mineração de dados ou para estudos de outra natureza, como estudos de associação ampla do genoma (*Genome-Wide Association Study* – GWAS). Tais estudos têm como objetivo analisar marcadores de um conjunto completo de DNA ou genoma de indivíduos da mesma espécie com o intuito de encontrar variações genéticas associadas a um determinado fenótipo (GHR, 2015a; NHGRI, 2015). Terceiro, a necessidade de reajuste e adaptação em aplicações nas quais as características dos padrões podem ser alteradas com o decorrer do tempo. E, por último, em análise exploratória de dados para descoberta de conhecimentos quanto à natureza dos dados, como eles se relacionam e sua estrutura (DUDA et al., 2001; MULDER et al., 2010; LINOFF; BERRY, 2011).

Dentro da análise de agrupamento, a validação é um processo importante que tem o propósito de avaliar a robustez dos resultados de agrupamento. Esse processo de validação apresenta técnicas que podem ser divididas em dois tipos: técnicas não-supervisionadas ou de validação interna e técnicas supervisionadas ou de validação externa. Sendo que técnicas de validação interna utilizam apenas informações que estão presentes na base de dados, limitação que não ocorre com as técnicas de validação externa (TAN et al., 2006; LIU et al., 2013).

Trabalhos recentes relacionados à validação interna de agrupamento (LIU et al., 2010, 2013) indicaram que algumas medidas de validação interna apresentam limitações. Ainda nesses trabalhos, são apresentadas orientações quanto à escolha de medidas de validação e uma nova medida de validação baseada em vizinhos mais próximos (*Clustering Valida-*

tion index based on Nearest Neighbors – CVNN), capaz de sugerir o número adequado de grupos e o melhor particionamento dos dados. Esses estudos têm como foco a análise, avaliação e desenvolvimento de medidas de validação interna, isto é, avaliação de resultados de agrupamento.

Ainda quanto à validação interna de agrupamento, Mulder et al. (2010) e Mulder (2014) introduziram uma abordagem que busca avaliar e melhorar os resultados de agrupamento por meio da identificação e remoção de amostras que não podem ser agrupadas adequadamente e por isso deveriam ser analisadas isoladamente. Dessa forma, tal abordagem além de avaliar os resultados de agrupamento também tem como foco melhorar os resultados por meio de validação e melhoria do conjunto de dados.

1.1 DEFINIÇÃO DO PROBLEMA

A abordagem de identificação e remoção de amostras introduzida por Mulder et al. (2010) e Mulder (2014) não analisa o comportamento – o percentual de amostras removidas e a instabilidade resultante – apresentado após a remoção de todas as amostras prejudiciais, analisando apenas a remoção de amostras individualmente. Sendo assim, essa abordagem sugere que a remoção de um subconjunto de amostras que foram avaliadas individualmente – quanto à preservação de estrutura – não prejudica a estrutura dos dados. Este fato não pode ser garantido.

De acordo com a dimensão da base de dados e a distribuição de suas amostras, a remoção de amostras isoladas pode não afetar a estrutura dos dados – que é aproximada pelo agrupamento médio – como discutido pelo conceito de preservação de estrutura dos dados (Seção 2.3.5). Porém, o fato de que essas amostras, isoladamente, não afetam a estrutura dos dados, não garante que o mesmo ocorra para o conjunto de todas essas amostras. Portanto, a remoção de amostras que apenas foram avaliadas isoladamente, pode afetar a estrutura dos dados, implicando assim em aumento de instabilidade.

Ainda considerando a forma pela qual as amostras são selecionadas, foi identificado que essa abordagem não apresenta nenhum tipo de tratamento para evitar remoção excessiva de amostras. Portanto, além de poder afetar a estrutura dos dados, implicando em aumento de instabilidade, a abordagem introduzida por Mulder et al. (2010) e Mulder (2014) também pode remover todo um grupo, o que pode ser considerado como indesejável de acordo com o contexto do estudo.

Considerando essas duas questões, o problema a ser tratado neste trabalho é a verificação do comportamento apresentado após a remoção de um subconjunto de amostras, com o intuito de evitar que a estrutura dos dados seja afetada; garantir redução de instabilidade e evitar remoção excessiva de amostras.

1.2 HIPÓTESE

Neste trabalho, tem-se como hipótese que por meio de ajustes no conceito de amostras prejudiciais aos resultados de agrupamento pode-se garantir aumento de estabilidade, evitar remoção excessiva de amostras, permitir que o usuário controle a análise e atribuir maior aplicabilidade e confiabilidade ao processo de análise.

1.3 OBJETIVOS

Considerando o problema e a hipótese apresentados anteriormente, o objetivo deste trabalho é atribuir maior aplicabilidade e confiabilidade em abordagens de avaliação e melhoramento de resultados de agrupamento que consideram avaliação e melhoramento do conjunto de dados. Esse objetivo geral pode ser divido em três objetivos específicos, a saber:

i) reformular o conceito de amostras prejudiciais que deveriam ser desconsideradas durante a tarefa de agrupamento e estudadas isoladamente;

ii) considerar a identificação dessas amostras como um problema de otimização no qual se tem como objetivo a redução da instabilidade do algoritmo de agrupamento sujeito à restrições quanto ao número de amostras para remoção;

iii) aquisição e validação de conhecimento sobre a natureza dos dados, seus relacionamentos e sua estrutura em base de dados genotípicos.

1.4 ORGANIZAÇÃO DO TEXTO

Este trabalho está organizado em cinco capítulos. No Capítulo 2 são apresentados conceitos fundamentais para compreensão deste trabalho, como análise de agrupamento e algoritmos genéticos. No Capítulo 3 é apresentada a metodologia utilizada, destacando-se

materiais, algoritmos, bases de dados e a abordagem proposta. No Capítulo 4 são apresentados os resultados obtidos por meio da aplicação da abordagem proposta e discussões para cada base de dados utilizada. Por fim, no Capítulo 5, são apresentadas as considerações finais, composto por uma avaliação geral da abordagem proposta considerando todos os ambientes de testes, pelas conclusões, contribuições, limitações e possibilidades de trabalhos futuros.

2 REFERENCIAL TEÓRICO

Neste capítulo são apresentados conceitos fundamentais para a compreensão deste trabalho, como análise de agrupamento e abordagens de validação e melhoramento de agrupamento. Ao final deste capítulo também é apresentado o algoritmo genético, algoritmo de busca utilizado para resolver o problema de otimização proposto neste trabalho (Seção 1.3). São apresentados também, brevemente, os algoritmos de agrupamento utilizados neste trabalho.

2.1 ANÁLISE DE AGRUPAMENTO

Análise de agrupamento é uma tarefa descritiva e não-supervisionada de mineração de dados que utiliza amostras não-rotuladas com o objetivo de encontrar grupos naturais, isto é, grupos de amostras fortemente relacionadas de forma que as amostras que pertençam a um mesmo grupo sejam mais similares entre si do que amostras em qualquer outro grupo (DUDA et al., 2001; TAN et al., 2006; LIU et al., 2013).

Existem diversas razões para a utilização de tarefas de aprendizado não-supervisionado, dentre elas podemos destacar algumas mais importantes. Inicialmente, a necessidade de rotular grandes bases de dados, o que poderia ser muito custoso sem o auxílio dessas tarefas. Segundo, sua utilização como uma tarefa preliminar para outras técnicas de mineração de dados ou para estudos de outra natureza, como estudos de associação ampla do genoma (*Genome-Wide Association Study* – GWAS). Terceiro, a necessidade de reajuste e adaptação em aplicações nas quais as características dos padrões podem ser alteradas com o decorrer do tempo. Por último, em análise exploratória de dados para descoberta de conhecimentos quanto à natureza dos dados, como eles se relacionam e sua estrutura (LANGLEY; SIMON, 1995; DUDA et al., 2001; SINGH et al., 2007; MULDER et al., 2010; LINOFF; BERRY, 2011).

Dentro da análise de agrupamento, a validação é um processo importante que tem o propósito de avaliar a robustez dos resultados de agrupamento. Esse processo apresenta técnicas que podem ser divididas em dois tipos: técnicas não-supervisionadas ou de validação interna e técnicas supervisionadas ou de validação externa. Sendo que técnicas de validação interna utilizam apenas informações que estão presentes na base de dados,

limitação que não ocorre para técnicas de validação externa (TAN et al., 2006; LIU et al., 2013). A Tabela 2.1 apresenta as medidas de validação interna utilizadas neste trabalho, destacando seus respectivos objetivos e valor ótimo.

Tabela 2.1: Classificação das medidas de validação interna utilizadas neste trabalho de acordo com seu objetivo. Para os índices de avaliação dos resultados de agrupamento, tem-se como objetivo maximizar seus valores. Para as medidas de avaliação e melhoramento dos resultados de agrupamento e do conjunto de dados, tem-se como objetivo minimizar seus valores.

Objetivo	Medidas	Valor ótimo
Avaliação dos resultados de agrupamento	Índice Calinski-Harabasz Índice de Dunn	max
Avaliação e melhoramento dos resultados de agrupamento por meio do melhoramento do conjunto de dados	Variância de estabilidade do agrupamento Instabilidade	min

2.1.1 ALGORITMO HDDC

O algoritmo HDDC (*High-Dimensional Data Clustering*) foi introduzido por Bouveyron et al. (2007) como um algoritmo de agrupamento que utiliza modelos de mistura gaussianas e o algoritmo EM (*Expectation-Maximization*) combinando ideias de agrupamento de subespaços para agrupamento de bases de dados de alta dimensionalidade, considerando que bases de dados de alta dimensionalidade geralmente são formadas por subespaços de baixa dimensionalidade escondidos no espaço original. Por sua vez, o algoritmo EM (DEMPSTER et al., 1977) é baseado em protótipo e também utiliza uma abordagem baseada em modelos estatísticos, mais especificamente em modelos de mistura nos quais são utilizadas distribuições estatísticas para definir grupos (TAN et al., 2006; MARSLAND, 2009).

O algoritmo HDDC disponibiliza catorze modelos gaussianos caracterizados e codificados de acordo com quatro critérios:

i) A_{kj}: os parâmetros dos subespaços dos grupos;

ii) B_k: os ruídos dos subespaços dos grupos;

iii) Q_k: a matriz de orientação de cada grupo;

iv) D_k: a dimensão intrínseca de cada grupo.

Cada critério é composto por diferentes tipos de restrições. O critério D, por exemplo, é composto por dois tipos de restrições: D_k e D. O tipo de restrição D_k permite que cada grupo tenha uma dimensão própria. Por outro lado, o tipo de restrição D restringe que a dimensão seja comum para todos os grupos. Dessa forma, cada modelo gaussiano é caracterizado e codificado de acordo com a combinação desses critérios e suas respectivas restrições. O modelo $A_{kj}B_kQ_kD_k$, por exemplo, é o modelo mais geral e, portanto, menos restritivo para os quatro critérios. Por outro lado, o modelo $ABQD$ é o modelo oposto, aquele mais específico e restritivo para os quatro critérios.

Esse algoritmo necessita basicamente de cinco parâmetros, o número de grupos esperados (K), o modelo gaussiano a ser utilizado (*model*), o número máximo de iterações (*itermax*), o critério de parada (*eps*) e o método de inicialização (*init*) (BOUVEYRON et al., 2007; BERGÉ et al., 2012).

A complexidade de tempo do algoritmo HDDC depende do modelo gaussiano (*model*) a ser utilizado e basicamente é definida como $O(Kpd)$, sendo K o número de grupos esperados, p a dimensão ou o número de atributos da base de dados e d a dimensão intrínseca dos grupos (BOUVEYRON et al., 2007; BERGÉ et al., 2012).

2.1.2 ALGORITMO SOM

O algoritmo SOM (*Self-Organizing Map* ou *Self-Organizing Feature Map*) foi introduzido por Kohonen (1990) como um tipo de Rede Neural Artificial baseado em aprendizado não-supervisionado e competitivo no qual os neurônios de saída da rede são organizados em forma de grade e competem entre si para pela ativação de cada amostra do conjunto de dados, sendo que tais amostras não são rotuladas (RUTKOWSKI, 2008; HAYKIN, 2009).

Geralmente, essa grade é bidimensional e cada neurônio é identificado por um par de coordenadas. Quando uma amostra do conjunto de dados é apresentada à camada de entrada da rede, a mesma é processada e atinge ou ativa um neurônio da grade. A localização de tal neurônio, o neurônio vencedor, é mantida e associada com a amostra que o ativou. Dessa forma, ao decorrer do processo de aprendizado, cada amostra pro-

cessada pela rede ajusta os neurônios de saída deformando a grade e, portanto, formando um mapa topográfico das amostras do conjunto de dados de modo que as coordenadas das deformações (neurônios vencedores) na grade indicam as característica intrínsecas do conjunto de dados e o relacionamento de suas amostras, isto é, as coordenadas de uma deformação na grade correspondem à projeção de uma característica do conjunto de dados como a formação de grupos, por exemplo. Esse fato explica a auto-organização da rede (KOHONEN, 1990; HAYKIN, 2009).

Esse algoritmo necessita basicamente de seis parâmetros, as dimensões da grade de neurônios de saída (*xdim* e *ydim*), o método de inicialização (*init*), a função utilizada para controle da taxa de aprendizado (*alphaType*), o tipo de vizinhança (*neigh*) e o tipo de topologia (*topol*) (KOHONEN et al., 1996; TAN et al., 2006; YAN, 2010).

2.1.3 ALGORITMO DBSCAN

O algoritmo DBSCAN (*Density-Based Spatial Clustering of Applications with Noises*) foi introduzido por Ester et al. (1996) como um algoritmo de agrupamento baseado em densidade desenvolvido para a descoberta de grupos com formatos arbitrários, para a identificação de ruídos e para apresentar eficiência em grandes bases de dados. Para algoritmos de agrupamento baseados em densidade, os grupos são regiões de alta densidade que estão separadas por regiões de baixa densidade (ESTER et al., 1996; TAN et al., 2006; WITTEN et al., 2011).

Esse algoritmo necessita de dois parâmetros, o raio que será considerado para a vizinhança (*Eps*) e o número mínimo de amostras ou pontos necessários em uma vizinhança para formar um grupo (*MinPts*). Geralmente, o parâmetro *Eps* é definido por um método heurístico também proposto por Ester et al. (1996), baseado na distância entre todos os pontos na base de dados e seus respectivos *MinPts* vizinhos mais próximos. Basicamente, a ideia é que o raio depende do número mínimo de pontos para formar um grupo e da distância entre todos os pontos na base de dados (ESTER et al., 1996; LIU et al., 2013).

A complexidade de tempo do algoritmo DBSCAN depende de n e t, sendo n o número de amostras do conjunto de dados e t o tempo necessário para encontrar a vizinhança de uma amostra. No pior caso, t é $O(n)$, caso no qual a complexidade de tempo do algoritmo é $O(n^2)$. Com o auxílio de uma estrutura de dados como R^*-*trees* para encontrar a vizinhança de uma amostra, é possível reduzir a complexidade de tempo do algoritmo

para $O(n \log n)$ (ESTER et al., 1996; TAN et al., 2006).

2.1.4 ALGORITMO K-*MEANS*

O algoritmo K-*means* foi introduzido por Lloyd (1982) como um algoritmo de agrupamento baseado em protótipo no qual é definido como um centroide, que geralmente é a média de um grupo de pontos ou um ponto central (BISHOP, 2006; TAN et al., 2006).

A complexidade de tempo do algoritmo K-*means* é $O(Kndi)$, sendo K o número de grupos esperados, n o número de amostras, d o número de atributos e i o número de iterações necessárias para o algoritmo convergir. Alguns autores resumem sua complexidade para $O(Kn)$, sendo K o número de grupos esperados e n o número de amostras(ISHIOKA, 2005; TAN et al., 2006).

2.2 VALIDAÇÃO INTERNA DE AGRUPAMENTO

Dentro das técnicas de validação de agrupamento, as medidas de validação interna geralmente são divididas em duas classes: medidas de coesão e medidas de separação. A primeira classe corresponde às medidas baseadas no critério de coesão ou compacidade. Esse critério tem como foco grupos isolados e portanto determina quão próximas ou quão relacionadas estão as amostras em determinado grupo, isto é, quão compacto e coeso é determinado grupo, sem considerar o relacionamento com os demais grupos encontrados. A segunda classe corresponde às medidas baseadas no critério de separação ou distinção. Esse critério tem como foco o relacionamento entre os grupos e por sua vez determina quão distinto ou bem separado um grupo está dos demais, isto é, quão isolado e distinto é determinado grupo (TAN et al., 2006).

Algumas medidas de validação interna consideram ambos os critérios, como o índice Calinski-Harabasz e os índices de Dunn (CALINSKI; HARABASZ, 1974; DUNN, 1974), enquanto outras medidas consideram apenas um critério, como a separação e os índices Γ (TAN et al., 2006; LIU et al., 2013). Independente de sua inspiração, essas medidas de validação tem como propósito a análise e validação dos resultados de agrupamento.

2.2.1 ÍNDICE CALINSKI-HARABASZ

O índice Calinski-Harabasz (CH) avalia os resultados de agrupamento com base na média entre a soma dos quadrados da distância entre grupos (separação) e da distância entre amostras dentro dos grupos (coesão). O índice CH é definido pela Equação 2.1, sendo n o número de amostras na base de dados, K o número de grupos, n_k o número de amostras dentro do grupo k, z_k o centroide do grupo k, c o centroide da base de dados e $d(x, y)$ a distância entre as amostras x e y. O número de grupos é definido como o valor K que maximiza o valor desse índice (CALINSKI; HARABASZ, 1974; MUFTI et al., 2005; LIU et al., 2010, 2013).

$$CH = \frac{\sum_{k=1}^{K} n_k d^2(z_k, c)/(K-1)}{\sum_{k=1}^{K} \sum_{i=1}^{n_k} d^2(x_i, z_k)/(n-K)} \tag{2.1}$$

2.2.2 ÍNDICE DE DUNN

O índice de Dunn (DI) avalia os resultados de agrupamento com base na menor distância entre amostras de diferentes grupos (separação) e o maior diâmetro dentro dos grupos (coesão). O índice DI é definido pela Equação 2.2, sendo K o número de grupos e C_i o i-ésimo grupo.

$$DI = \min_{1 \leq i \leq K} \left\{ \min_{1 \leq j \leq K, j \neq i} \left\{ \frac{dist(C_i, C_j)}{\max_{1 \leq k \leq K} \{diam(C_k)\}} \right\} \right\} \tag{2.2}$$

A menor distância entre amostras de diferentes grupos, $dist(C_i, C_j)$, é definida pela Equação 2.3 e o maior diâmetro dentro dos grupos, $diam(C_k)$, é definido pela Equação 2.4.

$$dist(C_i, C_j) = \min_{x \in C_i, y \in C_j} \{d(x, y)\} \tag{2.3}$$

$$diam(C_k) = \max_{x, y \in C_k} \{d(x, y)\} \tag{2.4}$$

Da mesma forma que o índice CH, o número de grupos é definido como o valor K que maximiza o valor desse índice (DUNN, 1974; HALKIDI et al., 2001; MAULIK; BANDYOPADHYAY, 2002; LIU et al., 2010, 2013).

2.3 MELHORAMENTO DO AGRUPAMENTO E DO CONJUNTO DE DADOS

Ainda dentro das técnicas de validação interna de agrupamento, robustez é um outro conceito importante que pode ser analisado com respeito à mudanças nas condições iniciais ou com respeito à mudanças na base de dados. Esse conceito difere do conceito das medidas de validação interna apresentado na Seção 2.2 pois além de validar os resultados de agrupamento, também tem como foco melhorar os resultados por meio de validação e melhoria no conjunto de dados.

Considerando esse conceito de robustez com respeito à mudanças na base de dados, Mulder et al. (2010) e Mulder (2014) introduziram duas medidas para robustez. A primeira medida, variância de estabilidade do agrupamento (*Cluster Stability Variance* – CSV), tem o propósito de avaliar a dependência quanto aos centroides apresentada pelo algoritmo de agrupamento. A segunda medida, instabilidade, tem o propósito de avaliar a estabilidade do relacionamento entre todas as amostras. Além dessas duas medidas, foram introduzidos dois conceitos para classificação de amostras. O primeiro conceito define como instáveis aquelas amostras que apresentam comportamento mais instável que o comportamento médio de todas as amostras. O segundo conceito define amostras que preservam a estrutura dos dados, isto é, amostras que quando removidas não afetam a estrutura dos dados e portanto também não afetam os resultados de agrupamento.

A partir dessas medidas e conceitos, Mulder et al. (2010) e Mulder (2014) desenvolveram uma abordagem de validação interna que busca avaliar e melhorar a estabilidade do algoritmo de agrupamento por meio de identificação e remoção de amostras que não podem ser agrupadas adequadamente (amostras instáveis) e por isso deveriam ser analisadas isoladamente (amostras que preservam a estrutura dos dados).

Nas próximas seções essas medidas de robustez e conceitos, utilizados para identificação e remoção de amostras, introduzidos por Mulder et al. (2010) e Mulder (2014), são brevemente apresentados com o intuito de serem autossuficientes para compreensão deste trabalho e sua proposta. Inicialmente, são apresentadas algumas definições essenciais para compreensão das medidas de robustez e dos conceitos utilizados para identificação e remoção de amostras. Uma vez que essas definições são apresentadas, as seções seguintes apresentam as medidas e os conceitos isoladamente.

2.3.1 DEFINIÇÕES PRELIMINARES

Dada uma base de dados $D = \{d_1, ..., d_n\}$ com n amostras, o i-ésimo de N resultados de agrupamento de um dado algoritmo de agrupamento A sobre D é representado como uma matriz $\boldsymbol{C}^i_{n\times n}$ na qual o elemento c^i_{jk} é definido pela Equação 2.5, sendo G_1 e G_2 dois grupos distintos do i-ésimo resultado de agrupamento, $1 \leq i \leq N$ e $1 \leq j, k \leq n$.

$$c^i_{jk} = \begin{cases} 1, & \text{se } d_j \in G_1,\ d_k \in G_2,\ G_1 \neq G_2 \\ 0, & \text{se } d_j,\ d_k \in G_1 \end{cases} \tag{2.5}$$

Considerando N resultados de agrupamento $\boldsymbol{C}^i_{n\times n}$, um conjunto de resultados de agrupamento M é definido pela Equação 2.6. A partir desse conjunto M, o agrupamento médio é representado como uma matriz $\overline{\boldsymbol{C}}_{n\times n}$ na qual o elemento $\overline{c}_{jk}$ é definido pela Equação 2.7.

$$M = \left\{\boldsymbol{C}^1_{n\times n}, ..., \boldsymbol{C}^N_{n\times n}\right\} \tag{2.6}$$

$$\overline{c}_{jk} = \frac{1}{N}\sum_{i=1}^{N} c^i_{jk} \qquad \text{para } 1 \leq j, k \leq n \tag{2.7}$$

2.3.2 VARIÂNCIA DE ESTABILIDADE DO AGRUPAMENTO

A variância de estabilidade do agrupamento (CSV) é a medida de robustez que tem o propósito de avaliar a dependência quanto aos centroides apresentada pelo algoritmo de agrupamento. Considerando que um agrupamento pode ser interpretado como uma variável aleatória, essa medida é uma aproximação para a variância de um agrupamento aleatório. Sendo assim, quanto menor a dependência, menor a medida CSV.

Dada uma base de dados $D = \{d_1, ..., d_n\}$ e um conjunto de resultados de agrupamento M (Equação 2.6) com agrupamento médio representado pela matriz $\overline{\boldsymbol{C}}_{n\times n}$ (Equação 2.7), a CSV de um algoritmo de agrupamento A sobre a base de dados D é definida pela Equação 2.8, sendo que a distância entre um resultado de agrupamento representado pela matriz $\boldsymbol{C}^i_{n\times n}$ e o agrupamento médio representado pela matriz $\overline{\boldsymbol{C}}_{n\times n}$, $d(\boldsymbol{C}^i, \overline{\boldsymbol{C}})$, é definida pela Equação 2.9.

$$CSV_D(A) = \frac{1}{2N}\sum_{i=1}^{N} d(\boldsymbol{C}^i, \overline{\boldsymbol{C}}) \tag{2.8}$$

$$d(\boldsymbol{C}^i, \overline{\boldsymbol{C}}) = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{j<k\leq n} |c^i_{jk} - \overline{c}_{jk}| \tag{2.9}$$

2.3.3 INSTABILIDADE

A instabilidade (μ) é a medida de robustez que tem o propósito de avaliar a estabilidade do relacionamento entre todas as amostras.

O elemento $\overline{c}_{jk}$ do agrupamento médio representado pela matriz $\overline{\boldsymbol{C}}_{n\times n}$ (Equação 2.7) representa a fração dos resultados de agrupamento no qual as amostras d_j e d_k estão em grupos diferentes. Sendo assim, menores valores indicam que essas amostras estão no mesmo grupo para a maioria dos casos, maiores valores indicam o oposto, e valores ao redor de 0,5 indicam que o relacionamento entre essas amostras é incerto. A partir desse relacionamento, uma função auxiliar de estabilidade (σ) é definida pela Equação 2.10, para $a \in [0, 1]$ (Figura 2.1).

$$\sigma(a) = \begin{cases} 1 - a, & \text{se } 0{,}5 \leq a \leq 1 \\ a, & \text{se } 0 \leq a < 0{,}5 \end{cases} \tag{2.10}$$

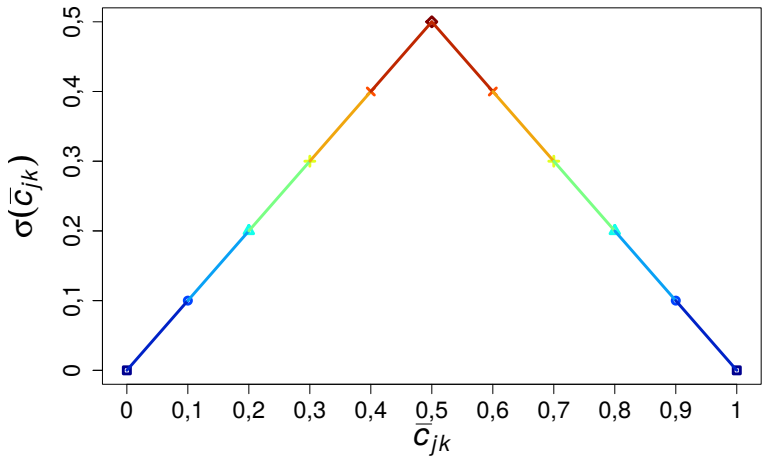

Figura 2.1: Comportamento da função auxiliar de estabilidade (σ). O elemento $\overline{c}_{jk}$ do agrupamento médio representa a fração dos resultados de agrupamento no qual as amostras d_j e d_k estão em grupos diferentes. Sendo assim, valores mais próximos de 0,5 indicam maior incerteza ou maior instabilidade quanto ao relacionamento entre tais amostras, como indicado por $\sigma(\overline{c}_{jk})$.

Dessa forma, $\sigma(\overline{c}_{jk})$ representa quão estável é a relação entre as amostras d_j and d_k, sendo que quanto menor esse valor mais estável o relacionamento. Uma vez que uma

função auxiliar de estabilidade (σ) foi definida, a instabilidade é uma medida que resume a estabilidade do relacionamento entre todas amostras, sendo que quanto maior o valor de $\sigma(a)$ maior a instabilidade e, portanto, maior a incerteza quanto às amostras que pertencem ao mesmo grupo (Figura 2.1).

Dada uma base de dados $D = \{d_1, ..., d_n\}$ com agrupamento médio representado pela matriz $\overline{\boldsymbol{C}}_{n \times n}$ (Equação 2.7), a instabilidade de uma amostra d_k com relação ao algoritmo de agrupamento A é definida pela Equação 2.11 e a instabilidade de um algoritmo de agrupamento A sobre a base de dados D é definida pela Equação 2.12.

$$\mu_D(d_k) = \frac{1}{n-1} \sum_{j \neq k}^{n} \sigma(\overline{c}_{jk}) \tag{2.11}$$

$$\begin{aligned} \mu_D(A) &= \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{j<k\leq n} \sigma(\overline{c}_{jk}) \\ &= \frac{1}{n} \sum_{k=1}^{n} \mu_D(d_k) \end{aligned} \tag{2.12}$$

2.3.4 RELAÇÃO ENTRE CSV E INSTABILIDADE

A partir das duas medidas de robustez, CSV e μ, alguns teoremas são apresentados e provados por Mulder et al. (2010) e Mulder (2014). O principal teorema afirma que dado um conjunto de resultados de agrupamento M (Equação 2.6), com agrupamento médio representado pela matriz $\overline{\boldsymbol{C}}_{n \times n}$ (Equação 2.7) e produzido por um algoritmo de agrupamento A, CSV é sempre menor ou igual à μ (Equação 2.13). Sendo assim, durante os testes apenas μ foi considerada.

$$CSV_D(A) \leq \mu_D(A) \tag{2.13}$$

2.3.5 AMOSTRAS INSTÁVEIS E PRESERVAÇÃO DA ESTRUTURA DOS DADOS

Amostras instáveis é o conceito utilizado para identificação das amostras que apresentam comportamento mais instável que o comportamento médio de todas as amostras da base de dados, isto é, amostras que são prejudiciais aos resultados de agrupamento. Dessa forma,

uma amostra d_k é considerada instável se a condição representada pela Equação 2.14 é verdadeira.

$$\mu_D(d_k) \geq \mu_D(A) \tag{2.14}$$

Desse conceito um outro teorema afirma que dada a amostra mais instável d_u na base de dados, isto é, $\mu_D(d_u) = \max_{1 \leq k \leq n} \{\mu_D(d_k)\}$, o agrupamento médio [1] após a remoção da amostra d_u, representado por $\overline{\boldsymbol{C}}_{D \setminus \{d_u\}}(D \setminus \{d_u\})$, implica que, no pior caso, a instabilidade resultante será igual a instabilidade original, enquanto no melhor caso, a instabilidade resultante será reduzida (Equação 2.15).

$$\mu_{D \setminus \{d_u\}}(A) \leq \mu_D(A) \tag{2.15}$$

Em seu primeiro trabalho, Mulder et al. (2010) afirmam que todas as amostras instáveis podem ser removidas, porém antes de removê-las, é necessário analisar outro conceito que é a estrutura dos dados. A estrutura dos dados diz respeito a forma pela qual um algoritmo de agrupamento enxerga a base de dados, forma que pode ser aproximada pelo agrupamento médio, sendo assim, cada algoritmo pode apresentar uma forma distinta. Dessa forma, apesar de ser considerada como instável, a remoção de uma amostra pode alterar a estrutura dos dados, o que é um comportamento indesejável que pode levar ao aumento da instabilidade.

Considerando esse conceito de estrutura dos dados, em seu estudo posterior Mulder (2014) introduziu o conceito de amostras que preservam a estrutura dos dados. Basicamente, uma amostra preserva a estrutura dos dados se a aplicação do algoritmo de agrupamento após sua remoção não afeta a estrutura que é aproximada pelo agrupamento médio.

Antes de definir o conceito de amostras que preservam a estrutura dos dados e apresentar um exemplo ilustrativo, é definida uma notação auxiliar que basicamente representa variações do agrupamento médio definido pela Equação 2.7 (pg. 28):

- $\overline{\boldsymbol{C}}_D(D)$ representa a matriz $\overline{\boldsymbol{C}}_{n \times n}$, correspondente ao agrupamento médio produzido pela aplicação de um algoritmo de agrupamento A sobre toda a base de dados D;

[1] O agrupamento médio após a remoção da amostras d_u também é definido de acordo com a Equação 2.7, porém é referenciado pela notação $\overline{\boldsymbol{C}}_{D \setminus \{d_u\}}(D \setminus \{d_u\})$ com o objetivo de destacar que tal agrupamento médio corresponde à aplicação do algoritmo de agrupamento A sobre a base de dados D desconsiderando a amostra d_u.

- $\overline{\boldsymbol{C}}_D(D \setminus \{d_p\})$ representa a matriz $\overline{\boldsymbol{C}}_{n \times n}$, correspondente ao agrupamento médio produzido pela aplicação de um algoritmo de agrupamento A sobre toda a base de dados D, desconsiderando a linha e a coluna relacionadas à amostra d_p. Ou seja, $\overline{\boldsymbol{C}}_D(D \setminus \{d_p\})$ representa uma matriz $m \times m$, sendo $m = n - 1$;

- $\overline{\boldsymbol{C}}_{D \setminus \{d_p\}}(D \setminus \{d_p\})$ representa a matriz $\overline{\boldsymbol{C}}_{(n-1)\times(n-1)}$, correspondente ao agrupamento médio produzido pela aplicação do mesmo algoritmo de agrupamento A sobre a base de dados D sem a amostra d_p. Ou seja, $\overline{\boldsymbol{C}}_{D \setminus \{d_p\}}(D \setminus \{d_p\})$ também representa uma matriz $m \times m$, sendo $m = n - 1$.

Considerando tal notação, dada uma base de dados $D = \{d_1, ..., d_n\}$, uma amostra d_p preserva a estrutura dos dados se a condição representada pela Equação 2.16 é verdadeira.

$$\overline{\boldsymbol{C}}_D(D \setminus \{d_p\}) = \overline{\boldsymbol{C}}_{D \setminus \{d_p\}}(D \setminus \{d_p\}) \tag{2.16}$$

A condição definida pela Equação 2.16 deve ser utilizada quando as mesmas condições iniciais são utilizadas antes e após a remoção de uma dada amostra, isto é, em casos nos quais os mesmos centroides são utilizados antes e após a remoção. Em casos nos quais diferentes centroides são utilizados, uma condição mais relaxada deveria ser considerada. Sendo assim, em casos nos quais diferentes centroides são utilizados, uma amostra d_p preserva a estrutura dos dados se a condição representada pela Equação 2.17 é verdadeira, para alguma norma $\|.\|$ e algum $\alpha > 0$, sendo α o fator que controla o quanto a condição deve ser relaxada.

$$\|\overline{\boldsymbol{C}}_D(D \setminus \{d_p\}) - \overline{\boldsymbol{C}}_{D \setminus \{d_p\}}(D \setminus \{d_p\})\| \leq \alpha \|\overline{\boldsymbol{C}}_D(D \setminus \{d_p\})\| \tag{2.17}$$

Com o intuito de ilustrar e facilitar a compreensão da notação utilizada e do conceito de preservação de estrutura dos dados, é apresentado um exemplo ilustrativo.

- *Exemplo ilustrativo:* Para exemplificar esse conceito de preservação de estrutura dos dados, considere uma base de dados $D = \{d_1, d_2, d_3, d_4, d_5\}$ e o agrupamento médio

$\overline{\boldsymbol{C}}_D(D)$ produzido por um algoritmo de agrupamento A e mostrado na Equação 2.18.

$$\overline{\boldsymbol{C}}_D(D) = \begin{bmatrix} 1 & 0{,}7 & \mathbf{0{,}2} & 0{,}3 & 0{,}8 \\ 0{,}7 & 1 & \mathbf{0{,}6} & 0{,}9 & 0{,}1 \\ \mathbf{0{,}2} & \mathbf{0{,}6} & \mathbf{1} & \mathbf{0{,}7} & \mathbf{0{,}9} \\ 0{,}3 & 0{,}9 & \mathbf{0{,}7} & 1 & 0{,}8 \\ 0{,}8 & 0{,}1 & \mathbf{0{,}9} & 0{,}8 & 1 \end{bmatrix} \tag{2.18}$$

Para verificar se a amostra d_3 preserva a estrutura dos dados, o algoritmo de agrupamento A é aplicado à $D \setminus \{d_3\}$ utilizando os mesmos centroides e produzindo um novo agrupamento médio $\overline{\boldsymbol{C}}_{D\setminus\{d_3\}}(D \setminus \{d_3\})$ mostrado pela Equação 2.19.

$$\overline{\boldsymbol{C}}_{D\setminus\{d_3\}}(D \setminus \{d_3\}) = \begin{bmatrix} 1 & 0{,}7 & \mathbf{0{,}4} & 0{,}8 \\ 0{,}7 & 1 & 0{,}9 & 0{,}1 \\ \mathbf{0{,}4} & 0{,}9 & 1 & 0{,}8 \\ 0{,}8 & 0{,}1 & 0{,}8 & 1 \end{bmatrix} \tag{2.19}$$

Dessa forma, a amostra d_3 não preserva a estrutura dos dados uma vez que a matriz mostrada pela Equação 2.18, sem a linha e coluna 3 é diferente da matriz mostrada pela Equação 2.19 (Equação 2.20).

$$\overline{\boldsymbol{C}}_D(D \setminus \{d_3\}) = \begin{bmatrix} 1 & 0{,}7 & \mathbf{0{,}3} & 0{,}8 \\ 0{,}7 & 1 & 0{,}9 & 0{,}1 \\ \mathbf{0{,}3} & 0{,}9 & 1 & 0{,}8 \\ 0{,}8 & 0{,}1 & 0{,}8 & 1 \end{bmatrix} \neq \overline{\boldsymbol{C}}_{D\setminus\{d_3\}}(D \setminus \{d_3\}) \tag{2.20}$$

A remoção da amostra d_3 afetou a forma pela qual o algoritmo de agrupamento enxergava o relacionamento entre as amostras d_1 e d_4, então d_3 não preserva a estrutura dos dados. Assim, para que a estrutura dos dados fosse preservada a remoção da amostra d_3 não deveria afetar o relacionamento entre as amostras d_1 e d_4, isto é, a matriz mostrada pela Equação 2.18, sem a linha e coluna 3, deveria ser igual à matriz mostrada pela Equação 2.19.

Considerando ambos os conceitos, de amostras instáveis e amostras que preservam a estrutura dos dados, Mulder (2014) afirma que todas as amostras instáveis que preservam

a estrutura dos dados podem ser removidas da base de dados, sem alterar a estrutura da base de dados e aumentando a robustez do algoritmo de agrupamento.

2.3.6 VISÃO GERAL DA ABORDAGEM

A partir dessas medidas e conceitos apresentados previamente, Mulder et al. (2010) e Mulder (2014) desenvolveram a abordagem de validação interna para identificação e remoção de amostras que não podem ser agrupadas adequadamente e portanto deveriam ser estudadas isoladamente.

Dada uma base de dados $D = \{d_1, ..., d_n\}$ e N resultados de agrupamento da aplicação de um algoritmo A sobre D, o primeiro passo da abordagem consiste da representação desses N resultados de agrupamento como N matrizes $\boldsymbol{C}^i_{n\times n}$ (Equação 2.5, pg. 28), a formação do conjunto desses resultados, o conjunto M (Equação 2.6), e a definição do agrupamento médio representado pela matriz $\overline{\boldsymbol{C}}_{n\times n}$ (Equação 2.7) desse conjunto M. Uma vez que o agrupamento médio $\overline{\boldsymbol{C}}_{n\times n}$ foi definido, o segundo passo consiste do cálculo das duas medidas de robustez, $CSV_D(A)$ (Equação 2.8, pg. 28) e $\mu_D(A)$ (Equação 2.12, pg. 30), para o algoritmo de agrupamento A sobre a base de dados D. Nesse momento, são conhecidos a dependência do algoritmo de agrupamento quanto aos centroides e a estabilidade do relacionamento geral entre as amostras da base de dados. Em seguida é calculada a instabilidade individual das amostras (Equação 2.11, pg. 30).

A partir do resultado de instabilidade $\mu_D(A)$ para o algoritmo de agrupamento e dos resultados de instabilidade $\mu_D(d_k)$ por amostra, são identificadas aquelas amostras consideradas instáveis (Equação 2.14, pg. 31), formando um conjunto U. Conhecendo esse conjunto, o próximo passo é verificar se a remoção dessas amostras afetam a estrutura dos dados. Nessa abordagem de Mulder (2014) a remoção dessas amostras é verificada individualmente, isto é, uma amostra instável $d_i \in U$ é retirada da base de dados, o algoritmo é aplicado sobre o restante da base $D \setminus \{d_i\}$ e então é verificado se a remoção dessa única amostra d_i afeta a estrutura dos dados (Equação 2.16 ou 2.17, pg. 32, de acordo com os critérios citados anteriormente quanto aos centroides). Após a verificação dessa amostra d_i, a mesma volta para a base de dados e uma nova amostra instável $d_j \in U$ é analisada.

Ao final desse processo, é conhecido um conjunto $S \subseteq U$ formado por amostras que teoricamente podem ser removidas sem alterar a estrutura dos dados enquanto aumentam

a estabilidade do algoritmo de agrupamento.

2.4 ALGORITMOS GENÉTICOS

Algoritmo genéticos (AG) são procedimentos de busca propostos por Holland (1975) com base na teoria de seleção natural de Darwin (1859) nos quais indivíduos de uma população são avaliados de acordo com uma função de avaliação ou aptidão com o intuito de identificar aqueles mais aptos para sobreviver e evoluir. Consequentemente, espera-se que a partir da reprodução de tais indivíduos, a cada nova geração suas características sejam mantidas para gerar novos indivíduos ainda mais aptos enquanto as características daqueles indivíduos menos aptos sejam descartadas, de modo que ao final do procedimento sejam encontrados aqueles indivíduos que apresentam avaliações mais próximas do valor ótimo esperado para tal função de aptidão (GOLDBERG, 1989; ARTERO, 2009).

Na prática, algoritmos genéticos são utilizados para resolver problemas de otimização com o objetivo de encontrar uma solução para maximizar ou minimizar determinada função. Para tanto, inicialmente é formada uma população ou um conjunto de indivíduos ou cromossomos que correspondem à possíveis soluções, sejam elas aleatórias ou pré-determinadas. Dessa forma, esses cromossomos que correspondem à possíveis soluções são compostos por genes, que por sua vez representam parâmetros para a função de aptidão. A partir dessa população é iniciado o processo de evolução, no qual cada cromossomo é avaliado de acordo com a função que se deseja otimizar – função de aptidão –, um percentual de cromossomos mais aptos são mantidos – seleção – e reproduzidos – cruzamento e mutação – formando uma nova geração. Esse processo de reprodução composto pela seleção, cruzamento e mutação é repetido até que determinado critério de parada seja atingido (ARTERO, 2009; SILVA, 2011).

Para utilização de algoritmos genéticos, inicialmente é definida a função de aptidão, isto é, a função que se deseja otimizar e o critério de parada. De acordo com a função de aptidão e seus parâmetros, é definida a regra de formação dos genes, ou seja, a representação dos componentes de uma possível solução, números reais ou binários, por exemplo. Definida a representação dos genes, são formados os cromossomos, combinando genes para formar possíveis soluções para o problema e por consequência é formada a população inicial. A partir dessa população, inicia-se o processo de evolução descrito anteriormente. Ao final é apresentado o cromossomo mais apto encontrado, que corresponde à melhor

solução encontrada para o problema (SILVA, 2011).

Dentre as técnicas de busca e otimização, os algoritmos genéticos apresentam características que os diferenciam de outras técnicas. A primeira característica é que a cada iteração ou geração, os algoritmos genéticos avaliam um subconjunto de possíveis soluções (população) e não apenas uma única solução (ARTERO, 2009). Ainda quanto à utilização de um subconjunto de possíveis soluções, outra característica dos algoritmos genéticos é que tal subconjunto é formado levando em consideração informação histórica quanto às soluções avaliadas anteriormente e não apenas de modo aleatório. Outras características são sua simplicidade, robustez mesmo em complexos espaços de busca, variabilidade de soluções avaliadas que evita ficar preso em ótimos locais e permite que encontre boas soluções mesmo partindo de pobres soluções iniciais (GOLDBERG, 1989; FIELDING, 2007; OLSON; DELEN, 2008).

3 METODOLOGIA

Neste capítulo é apresentada a metodologia utilizada para realização deste trabalho, como uma descrição geral da abordagem metodológica, detalhes de implementação e configuração dos algoritmos utilizados, características das bases de dados utilizadas e a abordagem proposta.

3.1 ABORDAGEM METODOLÓGICA

Considerando o problema, a hipótese e os objetivos apresentados anteriormente no Capítulo 1, foi desenvolvida uma abordagem metodológica que pode ser dividida em três passos:

i) *Agrupamento:* Inicialmente, para cada base de dados utilizada como teste foram definidos diferentes conjuntos de condições iniciais para cada algoritmo de agrupamento utilizado. Para o agrupamento da base de dados de SNPs foram utilizados os algoritmos HDDC, SOM e DBSCAN. O que justifica a utilização destes algoritmos é o fato de que eles são capazes de identificar o número de grupos da base de dados utilizando diferentes métodos. Essa capacidade de identificar o número de grupos permite a validação de conhecimento, enquanto as diferenças entre as abordagens atribuem maior confiabilidade aos resultados. Para agrupamento das bases de dados *Iris* e *Wine* foi utilizado o algoritmo K-*means*. O que justifica a utilização destas bases de dados e deste algoritmo é o fato de que ambos foram utilizados por Mulder (2014) durante a avaliação de sua abordagem. Sendo assim, neste trabalho a abordagem proposta é avaliada sob as mesmas condições utilizadas por Mulder (2014) e também é avaliada em outro cenário, especificamente, a base de dados de SNPs com o algoritmo DBSCAN.

ii) *Análise de estabilidade do agrupamento:* Após a aplicação dos algoritmos de agrupamento sobre as bases de dados, foram realizadas validações e análises de estabilidade dos resultados de agrupamento. Para a base de dados de SNPs foi realizada a validação dos resultados de agrupamento apresentados pelos algoritmos HDDC, SOM, e DBSCAN com o índice Calinski-Harabasz e o índice de Dunn, apresentados nas

Seções 2.2.1 e 2.2.2, pelas Equações 2.1 e 2.2 (pg. 26), respectivamente. Após a validação dos resultados de agrupamento com esses índices, foi realizada a análise de estabilidade dos resultados apresentados pelo algoritmo DBSCAN com as medidas de robustez, CSV e μ, apresentadas nas Seções 2.3.2 e 2.3.3, pelas Equações 2.8 e 2.12 (pg. 28 e 30), respectivamente. Devido ao custo computacional dos testes para análise de estabilidade, e considerando os resultados de agrupamento obtidos, apenas o algoritmo DBSCAN foi utilizado durante essa análise de estabilidade com as medidas de robustez introduzidas por Mulder (2014). Sendo assim, para obtenção e validação de resultados de agrupamento com medidas de validação interna, todos os três algoritmos foram utilizados, enquanto que para análise de estabilidade apenas o algoritmo DBSCAN foi utilizado. Uma vez que as bases de dados *Iris* e *Wine* foram utilizadas apenas com o intuito de comparar a abordagem proposta com a abordagem apresentada por Mulder (2014), e devido ao número de instâncias utilizadas, os resultados de agrupamento do algoritmo K-*means* não foram avaliados com as medidas de validação interna. Esses resultados de agrupamento foram utilizados para a análise de estabilidade com as medidas de robustez, CSV e μ.

iii) *Análise de estabilidade dos dados:* Nessas análises foram utilizados os conceitos – apresentados na Seção 2.3.5 – de amostras instáveis (Equação 2.14, pg. 31) e amostras que preservam a estrutura dos dados (Equações 2.16 e 2.17, pg. 32) introduzidos por Mulder et al. (2010) e Mulder (2014), juntamente com as adaptações sugeridas pela abordagem proposta neste trabalho (Seção 3.5). Para a base de dados de SNPs, agrupada pelo algoritmo DBSCAN, foi utilizada a Equação 2.16 (pg. 32) uma vez que o algoritmo DBSCAN utiliza uma abordagem baseada em densidade, na qual a alteração dos centroides não necessariamente afetaria os resultados de agrupamento. Por outro lado, para as bases de dados *Iris* e *Wine*, agrupadas pelo algoritmo K-*means*, foi utilizada a Equação 2.17 (pg. 32) uma vez que o algoritmo K-*means* depende dos centroides utilizados. O parâmetro α foi definido como 0,5 (mesmo valor utilizado por Mulder (2014) durante a avaliação de sua abordagem) e foi utilizada a norma $\|.\|_1$, que é definida como o máximo das somas absolutas das colunas de uma dada matriz [2].

[2]Dada uma matriz $\boldsymbol{M}_{l\times c}$ com elementos m_{ij}, a norma $\|.\|_1$ de tal matriz é definida como $\|\boldsymbol{M}\|_1 = \max_{1\leq j\leq c}\left\{\sum_{i=1}^{l}|m_{ij}|\right\}$ (BOLDRINI et al., 1980).

A Seção 3.3 apresenta mais detalhes quanto ao conjunto de condições iniciais e detalhes de implementação para cada algoritmo, enquanto a Seção 3.4 apresenta mais detalhes quanto às bases de dados, como suas características e tarefas de pré-processamento. A Seção 2.3.6 apresenta a abordagem de Mulder (2014) em detalhes, enquanto a Seção 3.5 apresenta a abordagem proposta neste trabalho, destacando suas diferenças e contribuições. Os resultados obtidos e discussões são apresentados no Capítulo 4, sendo que a Seção 4.1 apresenta os resultados de agrupamento, de validação e de análise de estabilidade para a base de dados de SNPs, a Seção 4.2 e a Seção 4.3 apresentam os resultados da análise de estabilidade para as base de dados *Iris* e *Wine* e, por último, a Seção 5.1 apresenta uma discussão geral dos resultados obtidos em todas as bases de dados.

3.2 MATERIAIS

Para implementação e testes, foi utilizado o *R*, ambiente de software e linguagem de programação para computação estatística (R Core Team, 2015). A Tabela 3.1 apresenta detalhes quanto aos pacotes utilizados.

Tabela 3.1: Pacotes utilizados para as tarefas de agrupamento e análise de estabilidade dos resultados de agrupamento e das bases de dados utilizadas neste trabalho (Tabelas 3.6, 3.7 e 3.8, pg. 45 e 46), respectivamente.

Pacote	Algoritmo utilizado	Referência
fpc	DBSCAN	(HENNIG, 2014)
HDclassif	HDDC	(BERGÉ et al., 2012)
som	SOM	(YAN, 2010)
stats	K-*means*	(R Core Team, 2015)
genalg	Algoritmo genético	(WILLIGHAGEN, 2014)

Para realização dos testes, foram utilizadas máquinas com arquitetura 64-*bits* com distribuição *Linux*, *Ubuntu* 13.10, *kernel* 3.11.0-26 kernel-generic, 32 *GB* de memória *RAM* e processador *Intel Xeon E5540*.

3.3 ALGORITMOS

Nesta seção são apresentadas as instâncias utilizadas para a tarefa de agrupamento, destacando para cada algoritmo: o pacote que foi utilizado para testes, o número de instâncias

construídas e suas respectivas configurações.

3.3.1 HDDC

Para realização dos testes com o algoritmo HDDC foi utilizado o pacote *HDclassif* (BERGÉ et al., 2012). Este pacote implementa a versão do algoritmo de Bouveyron et al. (2007) disponibilizando os catorze modelos gaussianos citados na Seção 2.1.1 e cinco métodos de inicialização.

Esta versão do algoritmo HDDC basicamente utiliza cinco parâmetros: o número de grupos esperados (K), o modelo gaussiano a ser utilizado (*model*), o número máximo de iterações (*itermax*), o critério de parada (*eps*) e o método de inicialização (*init*). O critério de parada (*eps*) e o número máximo de iterações (*itermax*) foram mantidos de acordo com os valores sugeridos pelo pacote *HDclassif*, 1×10^{-3} e 60, respectivamente.

Foi definido o intervalo $[2, ..., 4]$ para o número de grupos esperados (K) e foram utilizados três métodos de inicialização (*init*):

i) *kmeans*: os grupos são inicializados utilizando o algoritmo K-*means*;

ii) *mini-em*: os grupos são inicializados utilizando o algoritmo EM;

iii) *random*: os grupos são inicializados aleatoriamente.

Dentre os modelos gaussianos disponibilizados pelo pacote *HDclassif*, foi utilizado o modelo (*model*) $A_{kj}B_kQ_kD_k$. Este modelo foi escolhido por ser o menos restritivo – como citado na Seção 2.1.1 – e com o objetivo de não induzir o algoritmo definindo restrições quanto às características da base de dados. A Tabela 3.2 apresenta mais detalhes sobre a configuração das instâncias do algoritmo HDDC.

3.3.2 SOM

Para realização dos testes com o algoritmo SOM foi utilizado o pacote *som* (YAN, 2010). Este pacote implementa uma versão em *batch* do algoritmo de Kohonen et al. (1996) disponibilizando três métodos de inicialização do mapa, duas funções para controle da taxa de aprendizado e dois tipos de vizinhança e topologia.

Esta versão do algoritmo SOM utiliza basicamente seis parâmetros: as dimensões da grade de neurônios de saída (*xdim* e *ydim*), o método de inicialização (*init*), a função

Tabela 3.2: Parâmetros utilizados para as instâncias do algoritmo HDDC. Para todas as instâncias foram utilizados: o modelo gaussiano (*model*) $A_{kj}B_kQ_kD_k$, o número máximo (*itermax*) de 60 iterações e 1×10^{-3} como o critério de parada (*eps*). Dessa forma, a diferença entre as instâncias é o número de grupos (K) esperados e o método de inicialização (*init*).

Instância	K	*init*
1	2	
2	3	*kmeans*
3	4	
4	2	
5	3	*mini-em*
6	4	
7	2	
8	3	*random*
9	4	

utilizada para controle da taxa de aprendizado (*alphaType*), o tipo de vizinhança (*neigh*) e o tipo de topologia (*topol*). O tipo de vizinhança (*neigh*) e o tipo de topologia (*topol*) foram mantidos de acordo com os valores sugeridos pelo pacote *som*, vizinhança gaussiana (*gaussian*) e topologia hexagonal (*hexa*).

Para controle da taxa de aprendizado (*alphaType*) foram utilizadas as duas funções disponibilizadas: *linear* e *inverse*. Para inicialização (*init*) foram utilizados os três métodos disponibilizados:

i) *sample*: o mapa é inicializado utilizando amostras da base de dados;

ii) *random*: o mapa é inicializado aleatoriamente;

iii) *linear*: o mapa é inicializado utilizando grades lineares sobre os dois principais componentes.

A dimensão x (*xdim*) da grade de neurônios de saída foi definida como 3 e a dimensão y (*ydim*) foi definida como 2. Dessa forma, o algoritmo SOM foi aplicado sobre a base de dados de SNPs disponibilizando três neurônios de saída para os grupos conhecidos e outros três neurônios de folga para possíveis divisões entre os grupos da base de dados. Na Seção 3.4.1 são apresentados mais detalhes sobre essa base de dados. A Tabela 3.3 apresenta mais detalhes sobre a configuração das instâncias do algoritmo SOM.

Tabela 3.3: Parâmetros utilizados para as instâncias do algoritmo SOM. Para todas as instâncias foram utilizados: a dimensão x (*xdim*) como 3, a dimensão y (*ydim*) como 2, o tipo de vizinhança (*neigh*) gaussiana e a topologia (*topol*) hexagonal. Portanto, a diferença entre as instâncias é o método de inicialização (*init*) e a função utilizada para controle da taxa de aprendizado (*alphaType*).

Instância	*init*	*alphaType*
1	*sample*	*linear*
2		*inverse*
3	*random*	*linear*
4		*inverse*
5	*linear*	*linear*
6		*inverse*

3.3.3 DBSCAN

Para realização dos testes com o algoritmo DBSCAN foi utilizado o pacote *fpc* (HENNIG, 2014). Este pacote implementa uma versão do algoritmo DBSCAN capaz de utilizar dados representados como uma matriz de distância. Esta versão é relativamente mais rápida, porém apresenta maior custo em relação ao uso de memória. Utilizando essa versão, não é necessário calcular a matriz de distância dos dados a cada aplicação do algoritmo sobre a base de dados, o que justifica o ganho já mencionado, quanto ao custo de tempo.

O algoritmo DBSCAN, implementado no pacote *fpc* (HENNIG, 2014), basicamente necessita de três parâmetros: o número mínimo (*MinPts*) de amostras ou pontos necessários em uma vizinhança para a formação de um grupo, o raio (*Eps*) que será considerado para a vizinhança e o método (*method*) pelo qual os dados foram representados. Os parâmetros *MinPts* e *Eps* foram definidos de acordo com a heurística proposta por Ester et al. (1996) (e apresentada na Seção 2.1.3), na qual dado um valor para *MinPts*, é determinado um valor adequado para *Eps*. Sendo assim, foi definido um intervalo $[2, ..., 10]$ para *MinPts* e para cada valor nesse intervalo, a heurística foi utilizada para definir um valor adequado para o parâmetro *Eps*, resultando em nove instâncias. A Tabela 3.4 apresenta mais detalhes sobre a configuração das instâncias do algoritmo DBSCAN.

Tabela 3.4: Parâmetros utilizados para as instâncias do algoritmo DBSCAN. Para todas as instâncias foi utilizado o método (*method*) *dist* para representar os dados. O número mínimo de amostras em uma vizinhança (*MinPts*) e o raio considerado para tal vizinhança (*Eps*) foram definidos de acordo com a heurística proposta por Ester et al. (1996).

Instância	*MinPts*	*Eps*
1	2	155,3319
2	3	158,0601
3	4	160,4307
4	5	160,5335
5	6	160,6674
6	7	160,9814
7	8	161,7760
8	9	161,7869
9	10	161,9166

3.3.4 K-*MEANS*

Para a realização dos testes com o algoritmo K-*means* foi utilizado o pacote *stats* (R Core Team, 2015). Este pacote implementa quatro versões do algoritmo K-*means*, são elas: a versão de Hartigan e Wong (1979) que é a versão utilizada neste trabalho e apontada pela documentação do pacote *stats* como a melhor versão, a versão de MacQueen (1967), a versão de Lloyd (1982) e a versão de Forgy (1965).

Estas versões do algoritmo K-*means* basicamente utilizam três parâmetros: o número (K) de grupos esperados, o máximo (*iter.max*) de iterações e a versão (*algorithm*) que deve ser utilizada. Assim como em Mulder (2014), durante a avaliação de sua proposta, foram utilizadas 99 instâncias do algoritmo K-*means*, sendo que a escolha inicial dos centroides era aleatória – o que justifica a utilização desse número de instâncias – e essas instâncias foram distribuídas igualmente de acordo com o número (K) de grupos esperados (2, 3, ou 4 grupos). O parâmetro *iter.max* foi definido para 10 iterações e o parâmetro *algorithm* foi definido para a versão de Hartigan e Wong (1979), como sugerido pela documentação do pacote *stats*. A Tabela 3.5 apresenta as configurações das instâncias do algoritmo K-*means*.

Tabela 3.5: Parâmetros utilizados para as instâncias do algoritmo K-*means*. Para todas as instâncias foram utilizados: o número máximo (*iter.max*) de 10 iterações e a versão (*algorithm*) de Hartigan e Wong (1979). Dessa forma, a diferença entre as instâncias é o número de grupos esperados (K) e os centroides iniciais, uma vez que foi utilizada inicialização aleatória para tais centroides.

Instância	K
1	
⋮	2
33	
34	
⋮	3
66	
67	
⋮	4
99	

3.4 BASES DE DADOS

Nesta seção são apresentadas as três bases de dados utilizadas para avaliação da abordagem proposta – SNP, *Iris* e *Wine* –, destacando suas características, tarefas de pré-processamento e justificativas de utilização.

3.4.1 SNP

Polimorfismos de base única (*Single Nucleotide Polymorphisms* – SNP) são marcadores moleculares de variações na sequência de DNA em uma única base, isto é, a substituição de um nucleotídeo por outro, variações estas que são bastante comuns entre indivíduos da mesma espécie. Geralmente, essas variações são encontradas entre genes mas também podem ser encontradas dentro de genes ou regiões reguladoras, caso no qual essa variação pode afetar a função do gene causando diferenças entre os indivíduos. A análise de SNPs é considerada como um desafio computacional (MOUNT, 2004; GHR, 2015b).

Com o objetivo de adquirir e validar o conhecimento sobre a base de dados, e avaliar a abordagem proposta, uma base de dados de SNPs foi utilizada. A base de dados utilizada contém 2467 amostras de SNPs de três raças de gado bovino. Duas raças, Holandesa e Jersey, são taurinas e inicialmente apresentavam 56947 marcadores e a outra raça, Nelore,

é zebuína e inicialmente apresentava 54000 marcadores.

Devido à diferença quanto aos marcadores disponíveis em cada raça, foi necessário identificar os marcadores comuns entre as três. Foram encontrados 49725 marcadores comuns entre elas sendo que 136 deles eram constantes independentemente da raça, por isso foram desconsiderados durante a aplicação do algoritmo de agrupamento; foram utilizados, portanto, 49589 marcadores.

SNPs geralmente são representados de acordo com seu número de alelos A, dessa forma SNPs com alelos AA são representados com o número 2, SNPs com alelos AB ou BA são representados como 1, enquanto SNPs com alelos BB são representados como 0. Mais detalhes sobre a base de dados são apresentados na Tabela 3.6.

Tabela 3.6: Características da base de dados de SNPs. Esta base de dados é composta por SNPs de três raças de gado bovino: Holandesa, Jersey e Nelore.

Característica	Descrição
Grupos	3
Amostras	2467
Atributos	49589
Amostras Holandesa	577
Amostras Jersey	1024
Amostras Nelore	866

3.4.2 IRIS E WINE

Com o objetivo de avaliar a abordagem proposta sobre as mesmas condições utilizadas por Mulder (2014), duas bases de dados disponíveis no Repositório de Aprendizado de Máquina da Universidade da Califórnia, Irvine (*University of California, Irvine* – UCI) (LICHMAN, 2013) foram utilizadas. São elas:

i) *Iris*: Essa base de dados contém 150 amostras distribuídas igualmente entre três espécies de plantas do gênero íris, são elas: *Iris setosa*, *Iris versicolor* e *Iris virginica*. As amostras são compostas por atributos relacionados ao tamanho de sépala e pétala. Ver a Tabela 3.7 para mais detalhes sobre essa base de dados.

ii) *Wine*: Essa base de dados contém 178 amostras distribuídas entre três tipos de vinhos, sendo que todas as amostras foram cultivadas na mesma região, porém em

diferentes vinícolas. As amostras são compostas por atributos como o percentual de álcool, a intensidade da cor, entre outros. Para mais detalhes sobre essa base de dados, conferir a Tabela 3.8.

Tabela 3.7: Características da base de dados *Iris*. Esta base de dados é composta por amostras de três espécies de plantas do gênero íris: setosa, versicolor e virginica.

Características	Descrição
Grupos	3
Amostras	150
Atributos	4
Amostras setosa	50
Amostras versicolor	50
Amostras virginica	50

Tabela 3.8: Características da base de dados *Wine*. Esta base de dados é composta por amostras de três tipos de vinho: tipo 1, tipo 2 e tipo 3.

Características	Descrição
Grupos	3
Amostras	178
Atributos	13
Amostras tipo 1	59
Amostras tipo 2	71
Amostras tipo 3	48

3.5 ABORDAGEM PROPOSTA

Por meio de experimentos com a abordagem de identificação e remoção de amostras proposta por Mulder et al. (2010) e Mulder (2014), algumas características indesejáveis foram identificadas. A primeira característica consiste da forma pela qual as amostras são definidas como prejudiciais aos resultados de agrupamento. Nessa abordagem de Mulder (2014), as amostras candidatas para remoção são analisadas e identificadas individualmente, dessa forma o comportamento apresentado após a remoção de um subconjunto de

amostras não é analisado, o que leva à segunda característica, que consiste em casos nos quais a remoção dessas amostras não garante redução de instabilidade e, pelo contrário, pode implicar no aumento da instabilidade. Ainda quanto à forma pela qual as amostras são definidas como prejudiciais, a abordagem de Mulder (2014) não leva em consideração que pode marcar todo um grupo para remoção, o que também pode ser considerada uma característica indesejável de acordo com o contexto do estudo.

Considerando essas questões, a hipótese deste trabalho é que por meio de ajustes no conceito de amostras prejudiciais aos resultados de agrupamento e do uso de algoritmo genético para identificação de tais amostras, é possível:

i) garantir aumento de estabilidade;

ii) evitar remoção excessiva de amostras;

iii) permitir que o usuário controle a análise, atribuindo maior aplicabilidade e confiabilidade ao processo de análise.

Na abordagem proposta neste trabalho, o conceito de verificação de amostras que preservam a estrutura dos dados é estendido e a identificação de amostras para remoção é considerada como um problema de otimização em que o objetivo é a redução da instabilidade sujeito à restrições quanto ao número de amostras para remoção. Portanto, o algoritmo genético tem o propósito de encontrar um subconjunto de amostras instáveis que preserva a estrutura dos dados enquanto minimiza a instabilidade resultante e o número de amostras a serem removidas.

3.5.1 PRESERVAÇÃO DA ESTRUTURA DOS DADOS

Nesta abordagem não é considerada apenas a verificação de amostras – quanto à preservação da estrutura dos dados –, individualmente, mas também a verificação do comportamento apresentado após a remoção de um subconjunto de amostras. Essa verificação tem como propósito evitar que a remoção desse subconjunto afete a estrutura dos dados, o que poderia resultar em aumento de instabilidade.

Para isso, a mesma notação auxiliar apresentada na Seção 2.3.5, que basicamente representa variações do agrupamento médio definido pela Equação 2.7 (pg. 28), é estendida da seguinte forma:

- $\overline{\boldsymbol{C}}_D(D \setminus S)$ representa a matriz $\overline{\boldsymbol{C}}_{n \times n}$, correspondente ao agrupamento médio produzido pela aplicação de um algoritmo de agrupamento A sobre toda a base de dados D, desconsiderando as linhas e as colunas relacionadas às amostras do subconjunto $S \subseteq U \subseteq D$, sendo U o conjunto de amostras instáveis da base de dados D. Ou seja, $\overline{\boldsymbol{C}}_D(D \setminus S)$ representa uma matriz $m \times m$, sendo $m = n - |S|$;

- $\overline{\boldsymbol{C}}_{D \setminus S}(D \setminus S)$ representa uma matriz $\overline{\boldsymbol{C}}_{(n-|S|) \times (n-|S|)}$, correspondente ao agrupamento médio produzido pela aplicação do mesmo algoritmo de agrupamento A sobre a base de dados D sem as amostras do subconjunto $S \subseteq U \subseteq D$, sendo U o conjunto de amostras instáveis da base de dados D. Ou seja, $\overline{\boldsymbol{C}}_{D \setminus S}(D \setminus S)$ também representa uma matriz $m \times m$, sendo $m = n - |S|$.

Considerando tal notação, o conceito de verificação apresentado pela Equação 2.16 (pg. 32) é estendido para: dada uma base de dados $D = \{d_1, ..., d_n\}$, um subconjunto de amostras S preserva a estrutura dos dados se a condição representada pela Equação 3.1 é verdadeira.

$$\overline{\boldsymbol{C}}_D(D \setminus S) = \overline{\boldsymbol{C}}_{D \setminus S}(D \setminus S) \tag{3.1}$$

Da mesma forma, o conceito de verificação apresentado pela Equação 2.17 (pg. 32), casos nos quais diferentes centroides são utilizados, é estendido para: dada uma base de dados D, um subconjunto de amostras S preserva a estrutura dos dados se a condição representada pela Equação 3.2 é verdadeira para alguma norma $\|.\|$ e algum $\alpha > 0$.

$$\|\overline{\boldsymbol{C}}_D(D \setminus S) - \overline{\boldsymbol{C}}_{D \setminus S}(D \setminus S)\| \leq \alpha \|\overline{\boldsymbol{C}}_D(D \setminus S)\| \tag{3.2}$$

3.5.2 SELEÇÃO DAS AMOSTRAS PARA REMOÇÃO

Considerando estes ajustes quanto ao conceito de preservação da estrutura dos dados, dado um conjunto U de amostras instáveis que isoladamente preservam a estrutura dos dados, utilizou-se o algoritmo genético (AG) com o propósito de encontrar um subconjunto S de amostras instáveis que preserva a estrutura dos dados (Equação 3.1 ou 3.2, pg. 48) para minimizar a instabilidade resultante e o número de amostras removidas. Para tanto, o número de genes dos cromossomos foi definido como o número de amostras instáveis que isoladamente preservam a estrutura dos dados. Cada gene representa uma

amostra candidata à remoção, enquanto cada cromossomo representa um subconjunto dessas amostras candidato à remoção.

Com o propósito de fazer uma comparação, duas funções de avaliação foram desenvolvidas. Para a primeira função de avaliação do AG, $f(x)$, representada pela Equação 3.3, um cromossomo x é avaliado de acordo com a instabilidade apresentada pelo algoritmo de agrupamento A após a remoção do subconjunto S de amostras instáveis que isoladamente preservam a estrutura dos dados (subconjunto esse representado pelo cromossomo x).

$$f(x) = \mu_{D\setminus\{S(x)\}}(A) \tag{3.3}$$

Nessa primeira função de avaliação, a redução da instabilidade é o único fator considerado, de modo que o número de amostras removidas para encontrar a melhor instabilidade não foi considerado. Sendo assim, com essa função de avaliação, a abordagem proposta apenas tem o intuito de reduzir a instabilidade e encontrar um subconjunto de amostras que caso sejam removidas implicam nessa melhor instabilidade, independentemente do tamanho desse subconjunto.

Para a segunda função de avaliação do AG, $g(x)$, representada pela Equação 3.4, a primeira função de avaliação, $f(x)$, é estendida. Para $g(x)$, um cromossomo x é avaliado de acordo com a instabilidade apresentada pelo algoritmo de agrupamento A após a remoção do subconjunto S de amostras instáveis que isoladamente preservam a estrutura dos dados (subconjunto esse representado pelo cromossomo x), em relação ao tamanho do próprio conjunto S. Sendo $n = |S|$ o número de amostras do subconjunto S, $m = |x|$ o número de genes (ou o número total de amostras instáveis que isoladamente preservam a estrutura dos dados), isto é, o tamanho máximo de S e c um fator utilizado para controlar a relação entre a μ e n.

$$g(x) = \mu_{D\setminus\{S(x)\}}(A) \times \left(\frac{n}{m} + c\right) \tag{3.4}$$

Nessa segunda função de avaliação, a instabilidade é ponderada pelo número de amostras que deveriam ser removidas (n), de modo que a seleção de amostras para remoção é considerada como um problema de otimização com o propósito de encontrar um subconjunto de amostras que minimize a instabilidade resultante e, ao mesmo tempo, o número de amostras removidas. Como consequência, usando-se essa função de avaliação, espera-se

encontrar um subconjunto mínimo de amostras que, ao ser removido, implica em redução de instabilidade e evita-se a remoção excessiva de amostras, o que pode prevenir a remoção de todo um grupo.

Uma vez que a instabilidade resultante deve ser avaliada de acordo com o número de amostras que deveriam ser removidas, é necessário definir uma relação esperada entre essas duas medidas, isto é, definir em quanto seria necessário reduzir a instabilidade para se aumentar o número de amostras selecionadas para remoção. Na Equação 3.4 essa relação é representada pelo fator c. Dado um percentual p de redução esperada na instabilidade para que o número n de amostras selecionadas para remoção seja multiplicado pelo coeficiente k e o número total m de amostras instáveis que preservam a estrutura dos dados, o fator c é definido pela Equação 3.5, sendo $k > 1$, $0 < n \leq m$, $kn \leq m$ e $p \in]0,1[$. O comportamento do fator c considerando $n = 1$, $m = 25$, $p = \frac{1}{10}, \frac{2}{10}, ..., \frac{9}{10}$ e $k = 2, 3, ..., 10$ é ilustrado pela Figura 3.1.

$$c = \frac{n}{m}\left(\frac{k-1}{p} - k\right) \tag{3.5}$$

Analisando o comportamento do fator c, definido pela Equação 3.5 e ilustrado pela Figura 3.1, é possível notar que quanto maior o valor do coeficiente (k) para aumento de amostras removidas, maior o fator c, isto é, o fator c e o coeficiente k são proporcionais. Por outro lado, quanto menor o percentual (p) de redução esperada na instabilidade resultante, maior o fator c, isto é, c e p são inversamente proporcionais. Sendo assim, quanto menor o percentual p e quanto maior o coeficiente k, maior o fator c e, portanto, mais relaxada será a relação entre a instabilidade (μ) resultante e o número de amostras (n) removidas. Na prática, quanto mais relaxada esta relação entre μ e n, mais amostras podem ser selecionadas e removidas pois o coeficiente de aumento (k) é maior e está sob menor restrição quanto ao percentual (p) de redução esperada na instabilidade e, portanto, a redução da instabilidade tem maior prioridade. Por outro lado, quanto menos relaxada esta relação, menos amostras podem ser selecionadas e removidas pois o coeficiente (k) é menor e está sob maior restrição quanto ao percentual (p) de redução esperada na instabilidade e, portanto, a redução do número de amostras removidas tem maior prioridade.

Utilizando essa função de avaliação é possível combinar os dois objetivos, minimizar a instabilidade e minimizar o número de amostras a serem removidas. Em casos nos

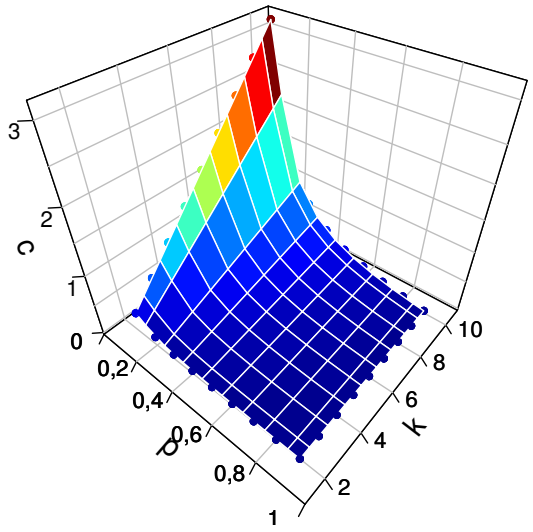

Figura 3.1: Comportamento do fator (c) que controla a relação entre instabilidade resultante e o número de amostras removidas, considerando uma amostra ($n = 1$), sendo 25 o número total de amostras ($m = 25$). O percentual de redução esperada na instabilidade está entre 10% e 90% ($p = \frac{1}{10}, \frac{2}{10}, ..., \frac{9}{10}$) e o coeficiente para aumento no número de amostras está entre 2 e 10 ($k = 2, 3, ..., 10$).

quais a instabilidade resultante é a mesma, prevalece aquele que apresenta menor número de amostras removidas, enquanto em casos nos quais o número de amostras removidas é o mesmo, prevalece aquele que apresenta menor instabilidade. Ainda sobre a função de avaliação $g(x)$ (Equação 3.4), é possível também definir qual a relação que deve ser considerada entre a instabilidade e o número de amostras para remoção por meio do fator c. Sendo assim, com essa função de avaliação, além das questões citadas anteriormente, a abordagem proposta também permite que o usuário tenha controle sobre o processo de análise, o que atribui maior aplicabilidade e confiabilidade à abordagem.

- *Exemplo ilustrativo:* Para exemplificar a segunda função de avaliação (Equação 3.4), considere que se deseja reduzir em 10% a instabilidade para que o número de amostras removidas aumente duas vezes (por exemplo, remover duas amostras ao invés de uma), sendo 25 o número máximo de amostras para remoção. Sendo assim, temos que $p = \frac{1}{10}$, $n = 1$, $k = 2$, $m = 25$, e portanto o fator c é representado pela

Equação 3.6, de acordo com a Equação 3.5.

$$c = \frac{1}{25}\left(\frac{2-1}{\frac{1}{10}} - 2\right) = \frac{8}{25} \tag{3.6}$$

Uma vez que o fator c foi definido, considere que determinado subconjunto S de amostras, composto por apenas uma amostra, resulta em 0,5 de instabilidade após sua remoção, ainda sendo 25 o número máximo de amostras para a remoção. Assim, $\mu_{D\setminus S}(A) = \frac{1}{2}$, $n = 1$, $m = 25$, $c = \frac{8}{25}$, e portanto a avaliação $g(x)$ desse subconjunto (ou cromossomo) é representada pela Equação 3.7, de acordo com a Equação 3.4.

$$g(x) = \frac{1}{2}\left(\frac{1}{25} + \frac{8}{25}\right) = \frac{9}{50} \tag{3.7}$$

Conhecida a avaliação $g(x)$ atribuída à esse cenário, considere um novo cenário no qual c, m e a avaliação se mantém, porém o número n de amostras removidas foi aumentado k vezes. Dessa forma, nesse cenário se está calculando a instabilidade que deve ser alcançada para que mesmo aumentando o número de amostras removidas em k vezes tenha-se a mesma avaliação apresentada pelo cenário no qual apenas uma amostra é removida. Como $n = 2$, $m = 25$, $c = \frac{8}{25}$, $g(x) = \frac{9}{50}$, a instabilidade que deve ser alcançada por esse subconjunto (ou cromossomo) é representada pela Equação 3.8, de acordo com a Equação 3.4.

$$\begin{aligned} \frac{9}{50} &= \mu_{D\setminus S}(A) \times \left(\frac{2}{25} + \frac{8}{25}\right) \\ \mu_{D\setminus S}(A) &= \frac{9}{20} \end{aligned} \tag{3.8}$$

Portanto, tem-se que um cromossomo x que apresente k vezes mais amostras selecionadas e removidas deve alcançar no mínimo uma redução de instabilidade equivalente à p para ser escolhido pelo algoritmo genético, isto é, a escolha desse cromos-

somo x está sujeita às restrições definidas na Expressão 3.9.

$$\begin{cases} g(x) = \mu_2\left(\frac{kn}{m}+c\right) \leq \mu_1\left(\frac{n}{m}+c\right) \\ k > 1 \\ 0 < n \leq m \\ kn \leq m \\ c = \frac{n}{m}\left(\frac{k-1}{p}-k\right) \\ p \in]0,1[\\ \mu_1 \in \left]0,\frac{1}{2}\right] \\ \mu_2 \leq \mu_1(1-p) \end{cases} \tag{3.9}$$

Para ambas as funções de avaliação, dado um cromossomo que representa um subconjunto de amostras instáveis que isoladamente preservam a estrutura dos dados, as instâncias de um dado algoritmo de agrupamento são aplicadas sobre a base de dados desconsiderando os genes ativos no cromossomo, isto é, o subconjunto de amostras. Dessa forma, esse subconjunto de amostras é avaliado considerando a instabilidade resultante e o número de amostras removidas de acordo com a Equação 3.3 ou a Equação 3.4.

Para a realização de testes com o algoritmo genético foi utilizado o pacote *genalg* (WILLIGHAGEN, 2014). Esse pacote apresenta uma versão do algoritmo genético que utiliza genes binários e basicamente necessita de sete parâmetros: o número de genes *size*, o tamanho da população *popSize*, o número de gerações *iters*, a chance de que um gene sofra mutação *mutationChance*, o número de genes que serão mantidos na próxima geração *elitism*, a taxa de genes ativos no cromossomo *zeroToOneRatio* e a função de avaliação *evalFunc*. A Tabela 3.9 sumariza os valores dos parâmetros utilizados pelo algoritmo genético.

Tabela 3.9: Parâmetros utilizados para o algoritmo genético.

Parâmetro	Descrição
Tipo de cromossomo	binário
Genes	o número total de amostras instáveis que isoladamente preservam a estrutura dos dados
Tamanho da população	20
Gerações	100
Chance de mutação	10%
Elitismo	20%
Taxa de genes ativos	25%

4 RESULTADOS E DISCUSSÃO

Neste capítulo são apresentados os resultados obtidos por meio da aplicação dos algoritmos de agrupamento e das abordagens para validação e melhoramento de dados e agrupamento descritas nas Seções 2.3 e 3.5. Para melhor apresentação e discussão, os resultados são apresentados em seções separadas para cada ambiente de testes, destacando resultados de agrupamento, análise de estabilidade do agrupamento e análise de estabilidade dos dados. Uma avaliação geral da abordagem, considerando todos os ambientes de testes, é apresentada no Capítulo 5.

4.1 SNP

A base de dados de SNPs foi utilizada com o intuito de aquisição e validação de conhecimento quanto à natureza dos dados, seus relacionamentos e sua estrutura, além de comparar a abordagem aqui proposta com a abordagem apresentada por Mulder (2014). Dessa forma, foram utilizados três algoritmos de agrupamento de diferentes abordagens (para mais detalhes sobre os algoritmos e as configurações das instâncias utilizadas, consultar Seções 2.1 e 3.3, respectivamente):

i) *HDDC*: algoritmo baseado em modelos de mistura gaussianas e no algoritmo EM e introduzido por Bouveyron et al. (2007);

ii) *SOM*: tipo de Rede Neural Artificial baseado em aprendizado não-supervisionado e competitivo e introduzido por Kohonen (1990);

iii) *DBSCAN*: algoritmo baseado em densidade para aplicações com ruídos e introduzido por Ester et al. (1996).

Após a aplicação desses algoritmos sobre a base de dados de SNPs, os resultados de agrupamento foram avaliados de acordo com o Índice CH e o Índice DI (apresentados nas Seções 2.2.1 e 2.2.2, respectivamente). Logo após, foi realizada a análise de estabilidade do agrupamento realizado pelo algoritmo DBSCAN.

Como citado anteriormente, essa base de dados é composta por 2467 amostras de SNPs de três raças de gado bovino: duas raças taurinas (Holandesa e Jersey) e uma zebuína

(Nelore). Considerando essa classificação, esperava-se que a separação entre as raças Holandesa e Jersey poderia apresentar maior dificuldade por ambas serem taurinas. Pelo mesmo motivo, esperava-se que a raça Nelore poderia apresentar maior facilidade para ser separada das demais. Logo, esperava-se que os resultados de agrupamento fossem compostos basicamente por quatro grupos, sendo um grupo isolado composto pelas amostras da raça Nelore, um segundo grupo formado pela junção de amostras das raças Holandesa e Jersey, um terceiro grupo para amostras mais distintas da raça Holandesa e, por último, um grupo para amostras mais distintas da raça Jersey (Figura 4.1).

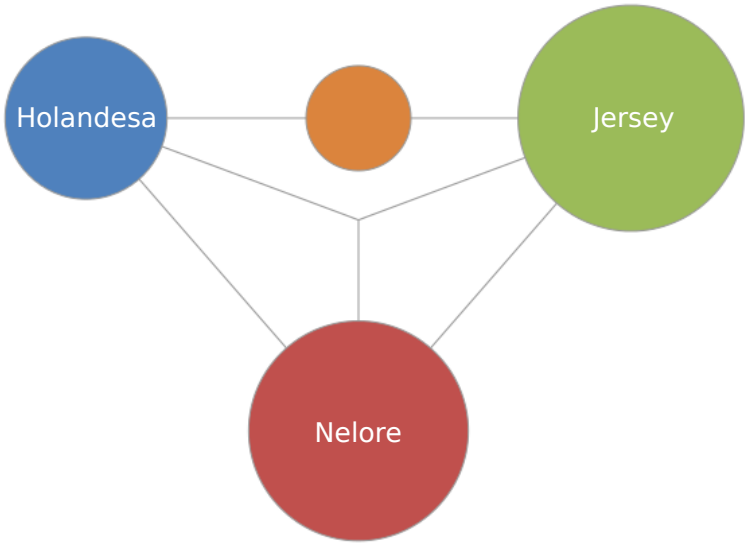

Figura 4.1: Resultado de agrupamento esperado para a base de dados de SNPs. O resultado de agrupamento esperado é composto por quatro grupos, sendo um grupo isolado composto pelas amostras da raça Nelore (zebuína), um segundo grupo formado pela junção de amostras das raças Holandesa e Jersey (taurinas), um terceiro grupo para amostras mais distintas da raça Holandesa e, por último, um grupo para amostras mais distintas da raça Jersey.

4.1.1 ALGORITMO HDDC

Os resultados de agrupamento obtidos pela aplicação das instâncias do algoritmo HDDC (Tabela 3.2, pg. 41) são ilustrados pela Figura 4.2 e suas respectivas avaliações são apresentadas e ilustradas pela Tabela 4.1 e Figura 4.3.

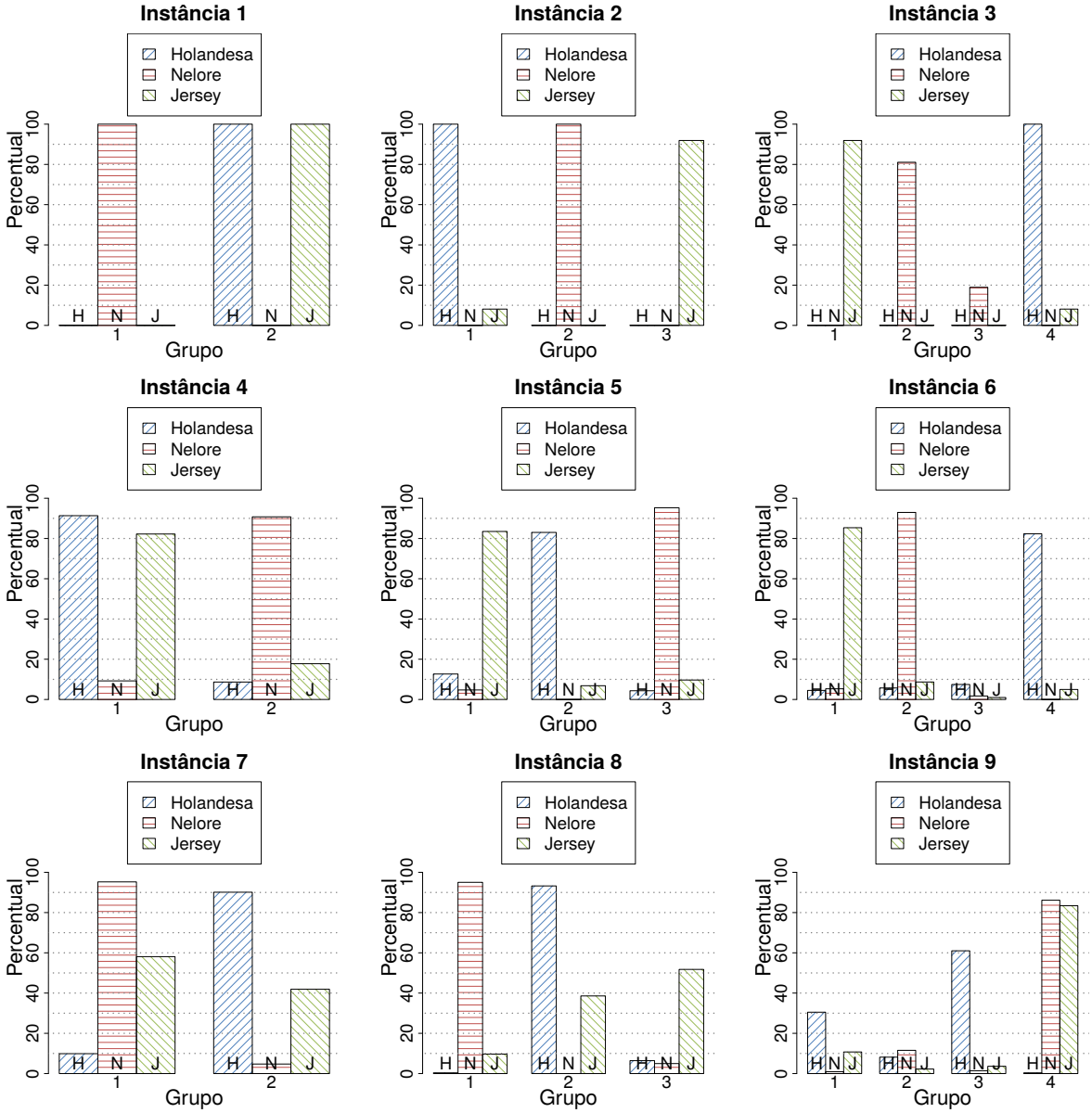

Figura 4.2: Resultados de agrupamento das instâncias do algoritmo HDDC sobre a base de dados de SNPs.

Analisando os resultados de agrupamento apresentados pelas instâncias do algoritmo HDDC (Figura 4.2), é possível notar que apenas três instâncias (instâncias 1, 2 e 3, inicializadas com o algoritmo K-*means* e $K = 2, 3, 4$, respectivamente) conseguiram formar grupos exclusivos para amostras da raça Nelore. A instância (1) formou dois grupos, sendo o primeiro composto pelas amostras da raça Nelore e o segundo pelas amostras das raças Holandesa e Jersey. As instâncias (2) e (3) também apresentam o comportamento esperado, amostras Nelore em grupo exclusivo e grupos compostos por amostras Holandesa e Jersey, porém é possível notar que todas as amostras da raça Holandesa estão concentradas em um único grupo compartilhado com aproximadamente 10% das amostras da raça

Jersey, ou seja, para essas instâncias é possível notar grupos consideravelmente dedicados para as raças Holandesa e Jersey. Uma outra característica interessante apresentada pelos resultados da instância (3) é a separação das amostras da raça Nelore em dois grupos. Esse fato pode ser um indicativo de que existe uma subdivisão dentro da raça Nelore, como se algum marcador ou um conjunto de marcadores (características) distinguissem tais amostras mesmo sendo da mesma raça.

Considerando os demais resultados de agrupamento do algoritmo HDDC, ainda é possível notar a formação de grupos dedicados à grandes percentuais de uma única raça (instâncias 5 e 6, por exemplo). É possível notar que alguns grupos não são claros (grupo 4, instância 9) por apresentarem altos percentuais para raças distintas. De modo geral, também é possível notar que amostras da raça Nelore foram distribuídas entre grupos dedicados às amostras das raças Holandesa e Jersey, não apresentando grupos exclusivos, portanto.

Tabela 4.1: Validação dos resultados de agrupamento do algoritmo HDDC para a base de dados de SNPs de acordo com o Índice Calinski-Harabasz (CH) e o Índice de Dunn (DI). A primeira instância do algoritmo HDDC maximiza os valores dos índices CH e DI e, portanto, aponta que dois grupos seriam a formação mais adequada.

Instância	Grupos	CH	DI
1	2	**1200,6210**	**1,2970**
2	3	882,0109	1,1499
3	4	593,7111	0,7717
4	2	579,0716	1,1738
5	3	559,1177	1,0911
6	4	380,2176	0,9967
7	2	269,9035	1,1164
8	3	504,8362	1,0155
9	4	87,4265	0,9844

Considerando a avaliação de tais resultados de agrupamento (Tabela 4.1 e Figura 4.3), é possível notar que houve concordância entre as medidas de validação interna, Índice CH e Índice DI. Ambos os índices apontaram a instância (1) como aquela que apresenta o melhor resultado de agrupamento e que apresenta o número de grupos mais adequado. Considerando tal avaliação, o melhor resultado é aquele que isola as amostras da raça Nelore e agrupa todas as amostras das raças Holandesa e Jersey em um único grupo,

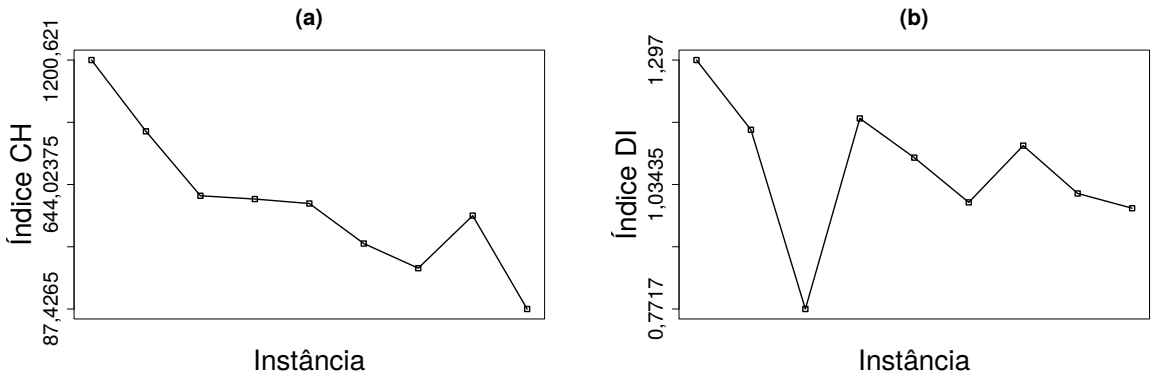

Figura 4.3: Validação dos resultados de agrupamento do algoritmo HDDC para a base de dados de SNPs de acordo com o Índice Calinski-Harabasz (CH) (a) e o Índice de Dunn (DI) (b). A primeira instância do algoritmo HDDC maximiza os valores de ambos os índices.

isto é, o melhor resultado é aquele que separa a raça zebuína (Nelore) das raças taurinas (Holandesa e Jersey). Como citado anteriormente, esperava-se a formação de grupos compostos por amostras das raças Holandesa e Jersey porém também esperava-se que houvessem grupos dedicados para cada uma dessas raças, destacando assim características que são distintas mesmo ambas sendo taurinas. Uma questão que pode ser notada é a influência da utilização do algoritmo K-*means* como um método de inicialização para a avaliação do Índice CH. As três instâncias que foram inicializadas com o algoritmo K-*means* (instâncias 1, 2 e 3) são aquelas que apresentam melhores avaliações de acordo com o Índice CH.

4.1.2 ALGORITMO SOM

Os resultados de agrupamento obtidos pela aplicação das instâncias do algoritmo SOM (Tabela 3.3, pg. 42) são ilustrados pela Figura 4.4 e suas respectivas avaliações são apresentadas e ilustradas pela Tabela 4.2 e Figura 4.5.

Analisando os resultados de agrupamento apresentados pelas instâncias do algoritmo SOM (Figura 4.4), é possível notar que apesar de seis instâncias, apenas dois resultados distintos de agrupamento foram apresentados (instâncias 1, 3 e 5 ou instâncias 2, 4 e 6) e ambos apresentam grupos exclusivos para as amostras da raça Nelore. Esse fato pode ser um indicativo de que as amostras da raça Nelore apresentam um conjunto de marcadores (características) muito distinto do conjunto de marcadores das amostras das raças taurinas, Holandesa e Jersey. As instâncias (1), (3) e (5) (utilizando a função *linear*

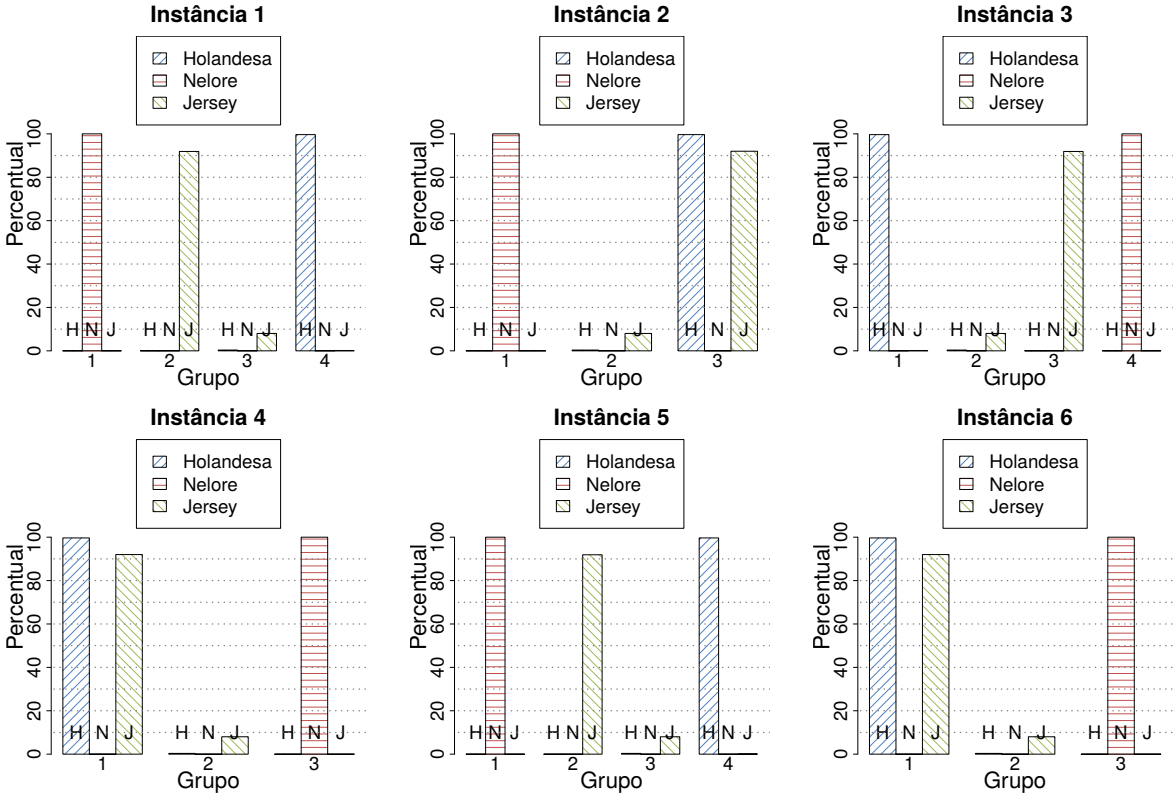

Figura 4.4: Resultados de agrupamento das instâncias do algoritmo SOM sobre a base de dados de SNPs.

para controle da taxa de aprendizado) apresentam a formação de quatro grupos, sendo um grupo exclusivo para as amostras da raça Nelore, um segundo grupo dedicado para as amostras da raça Jersey, um terceiro grupo dedicado para as amostras da raça Holandesa e um quarto grupo composto pela junção de aproximadamente 8% de amostras da raça Jersey e menos de 1% de amostras da raça Holandesa.

Basicamente a diferença é que as instâncias (2), (4) e (6) (utilizando a função *inverse* para controle da taxa de aprendizado) não apresentam grupos dedicados para as amostras das raças Holandesa e Jersey, formando assim, um único grupo composto por aproximadamente 92% de amostras da raça Jersey e 100% das amostras da raça Holandesa.

Considerando a avaliação de tais resultados de agrupamento (Tabela 4.2 e Figura 4.5), é possível notar que não houve concordância entre as medidas de validação interna, Índice CH e Índice DI. O Índice CH aponta as instâncias (2), (4) e (6) como aquelas que apresentam os melhores resultados de agrupamento e, portanto, o número de grupos mais adequado. Por outro lado, o Índice DI aponta que as instâncias (1), (3) e (5) apresentam os melhores resultados de agrupamento. Considerando o Índice CH, assim como para o algoritmo HDDC, o melhor resultado é aquele que isola as amostras da raça

Tabela 4.2: Validação dos resultados de agrupamento do algoritmo SOM para a base de dados de SNPs de acordo com o Índice Calinski-Harabasz (CH) e o Índice de Dunn (DI). As instâncias (2), (4) e (6) do algoritmo SOM maximizam os valores do índice CH e as instâncias (1), (3) e (5) maximizam os valores do índice DI.

Instância	Grupos	CH	DI
1	4	648,1619	**1,0206**
2	3	**674,3099**	1,0074
3	4	648,1619	**1,0206**
4	3	**674,3099**	1,0074
5	4	648,1619	**1,0206**
6	3	**674,3099**	1,0074

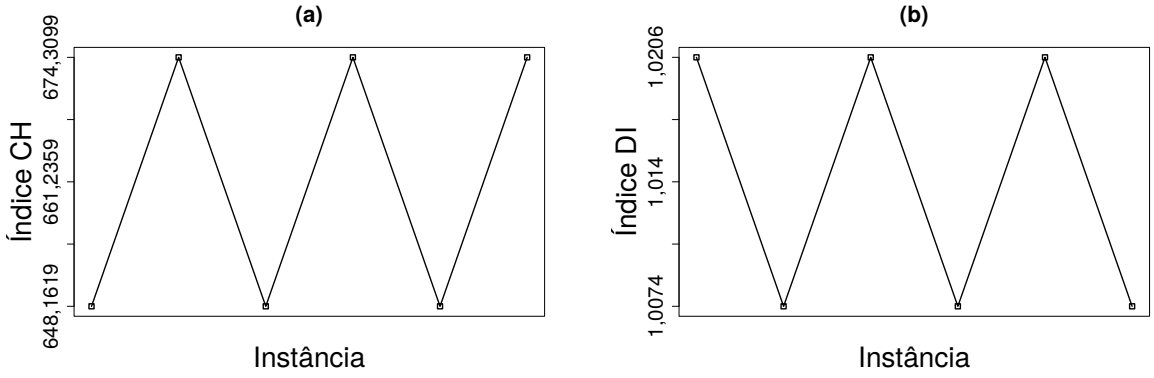

Figura 4.5: Validação dos resultados de agrupamento do algoritmo SOM para a base de dados de SNPs de acordo com o Índice Calinski-Harabasz (CH) (a) e o Índice de Dunn (DI) (b). As instâncias (2), (4) e (6) do algoritmo SOM maximizam os valores do índice CH e as instâncias (1), (3) e (5) maximizam os valores do índice DI.

Nelore e agrupa aproximadamente 92% e 100% das amostras das raças Jersey e Holandesa, respectivamente, em um único grupo, isto é, o melhor resultado é aquele que praticamente separa a raça zebuína (Nelore) das raças taurinas (Holandesa e Jersey). Considerando o Índice DI, o melhor resultado é aquele que isola as amostras da raça Nelore, forma grupos dedicados para as amostras das raças Holandesa e Jersey e um quarto grupo formado pela junção de baixos percentuais de amostras das raças taurinas. Outra questão que pode ser notada é a influência da função utilizada para controle da taxa de aprendizado na formação dos grupos e nos resultados da avaliação com os Índices CH e DI. Utilizando a função *linear* foram formados quatro grupos (instâncias 1, 3 e 5) que maximizam o Índice DI e utilizando a função *inverse* foram formados três grupos (instâncias 2, 4 e 6) que maximizam o Índice CH.

4.1.3 ALGORITMO DBSCAN

Os resultados de agrupamento obtidos pela aplicação das instâncias do algoritmo DBSCAN (Tabela 3.4, pg. 43) são ilustrados pela Figura 4.6 e suas respectivas avaliações são apresentadas e ilustradas pela Tabela 4.3 e Figura 4.7.

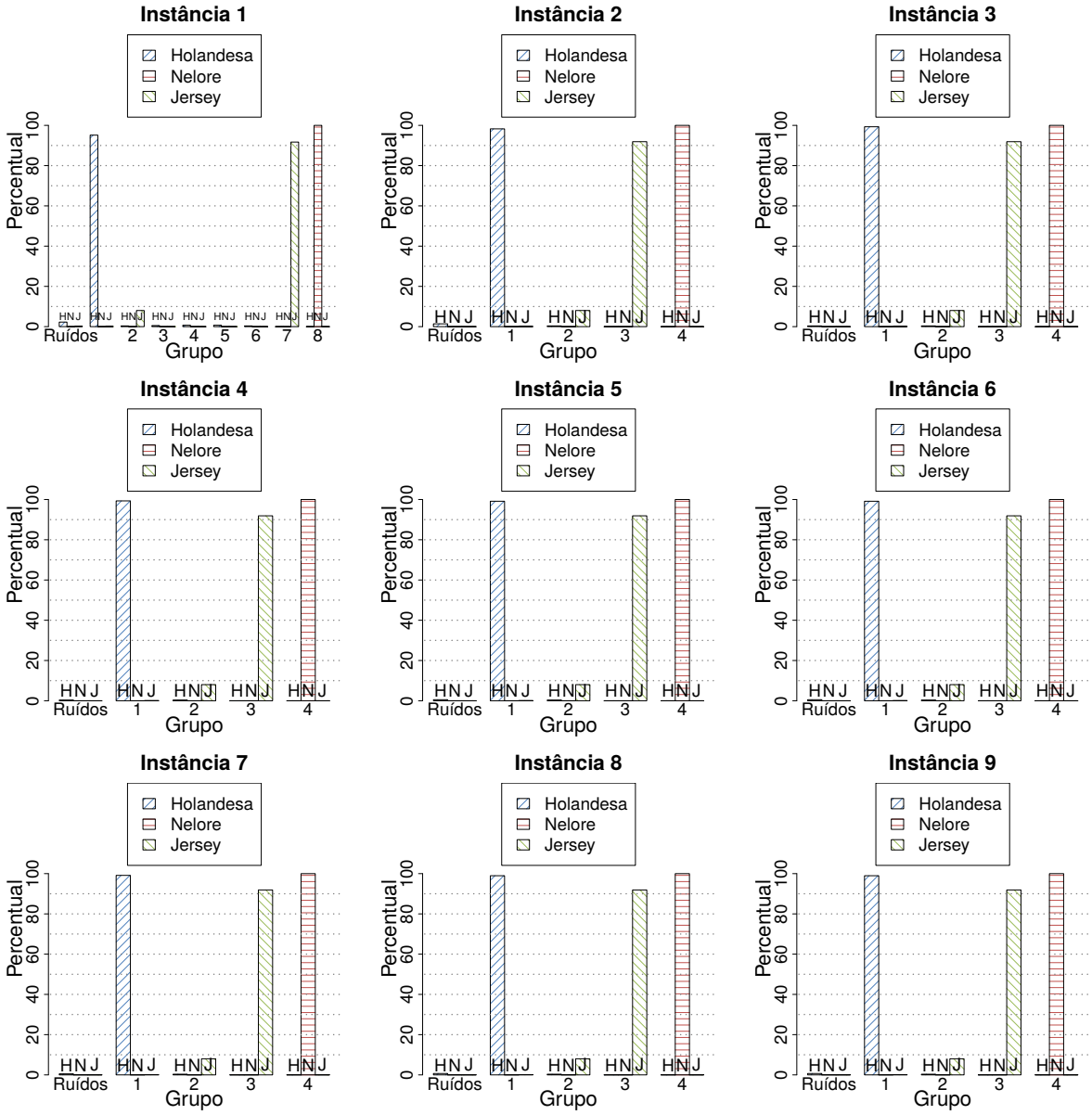

Figura 4.6: Resultados de agrupamento das instâncias do algoritmo DBSCAN sobre a base de dados de SNPs.

Analisando os resultados de agrupamento apresentados pelas instâncias do algoritmo DBSCAN (Figura 4.6), é possível notar que todas as instâncias apresentaram resultados muito próximos. A instância (1) apresentou oito grupos que ainda assim são muito próximos dos demais uma vez que a diferença entre seus grupos é a formação de quatro grupos

(3, 4, 5 e 6) compostos por menos que 1% das amostras da raça Holandesa.

Também é possível notar que todas as instâncias formaram grupos dedicados para as amostras das raças Holandesa e Jersey, grupos exclusivos para as amostras da raça Nelore e um grupo composto por aproximadamente 8% das amostras da raça Jersey (grupo 2 para todas as instâncias). Tal fato pode ser outro indicativo de que as amostras da raça Nelore apresentam um conjunto de marcadores (características) muito distinto do conjunto de marcadores das amostras das raças taurinas, Holandesa e Jersey, e um indicativo de que existe uma subdivisão dentro da raça Jersey.

O baixo percentual de ruídos apresentado é outra semelhança entre os resultados de todas as instâncias e pode ser um indicativo de confiabilidade para o processo de obtenção do conjunto de dados. No pior caso, a instância (1) aponta aproximadamente 2% das amostras da raça Holandesa como ruídos.

Tabela 4.3: Validação dos resultados de agrupamento do algoritmo DBSCAN para a base de dados de SNPs de acordo com o Índice Calinski-Harabasz (CH) e o Índice de Dunn (DI). As instâncias (3) e (4) do algoritmo DBSCAN maximizam os valores dos índices CH e DI.

Instância	Grupos	Ruídos	CH	DI
1	8	15	222,9343	0,8728
2	4	8	539,4474	0,9326
3	4	2	**581,4065**	**0,9751**
4	4	2	**581,4065**	**0,9751**
5	4	3	566,6536	0,9542
6	4	3	566,6536	0,9542
7	4	3	566,6536	0,9542
8	4	4	554,7669	0,9384
9	4	4	554,7669	0,9384

Considerando a avaliação de tais resultados de agrupamento (Tabela 4.3 e Figura 4.7), é possível notar que houve concordância entre as medidas de validação interna, Índice CH e Índice DI. Ambos os índices apontaram as instâncias (3) e (4) como aquelas que apresentam os melhores resultados de agrupamento e que apresentam o número de grupos mais adequado, sendo que, basicamente, as diferenças entre tais instâncias para as demais são as amostras que compõem cada grupo. Sendo assim, considerando tal avaliação, o melhor resultado é aquele que isola as amostras da raça Nelore, forma grupos dedicados

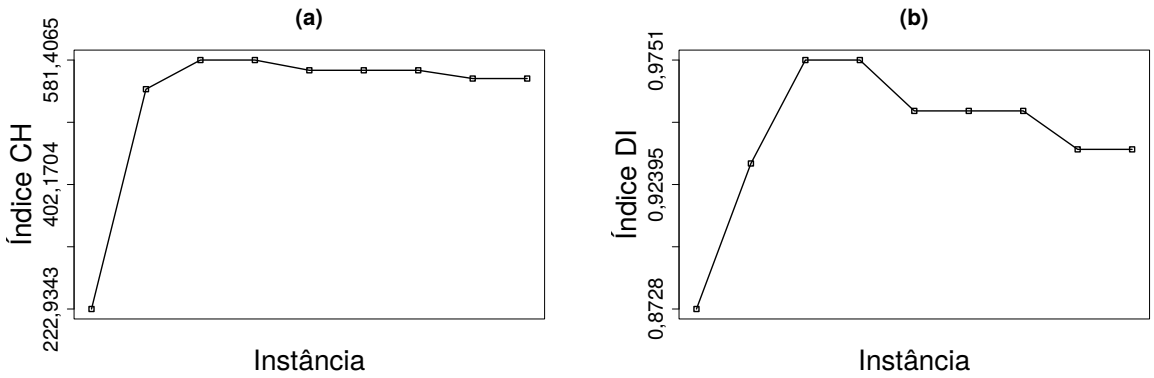

Figura 4.7: Validação dos resultados de agrupamento do algoritmo DBSCAN para a base de dados de SNPs de acordo com o Índice Calinski-Harabasz (CH) (a) e o Índice de Dunn (DI) (b). As instâncias (3) e (4) do algoritmo DBSCAN maximizam os valores de ambos os índices.

para as amostras das raças Holandesa e Jersey e um quarto grupo formado pela junção de baixos percentuais de amostras das raças taurinas. Uma questão que pode ser notada é que à partir da instância (4), a avaliação do Índice CH se estabiliza e a avaliação do Índice DI não apresenta comportamento tão estável quanto o comportamento do Índice CH mas apresenta uma queda moderada. Outra questão é a avaliação da instância (1) para ambos os índices. Considerando os resultados de agrupamento ilustrados pela Figura 4.6, não é possível notar grande diferença para os resultados de agrupamento das demais instâncias, porém analisando a Figura 4.7 pode-se notar que o agrupamento dessa instância é avaliado como consideravelmente inferior aos demais.

4.1.4 CSV E INSTABILIDADE

A partir dos resultados de agrupamento da base de dados de SNPs apresentados pelo algoritmo DBSCAN foi possível realizar a análise de estabilidade do agrupamento. Os resultados dessa análise indicaram $\mu_D(A) = 7{,}28 \times 10^{-4}$ para a instabilidade e $CSV_D(A) = 6{,}14 \times 10^{-4}$ para a variância de estabilidade do agrupamento. Após a análise de estabilidade de agrupamento foram realizadas análises de estabilidade dos dados.

4.1.5 SELEÇÃO E REMOÇÃO DE AMOSTRAS PREJUDICIAIS

Como citado no exemplo ilustrativo apresentado na Seção 2.3.5, as amostras instáveis que preservam a estrutura dos dados são identificadas por meio de uma nova aplicação do algoritmo de agrupamento sobre a base de dados desconsiderando-se tais amostras. O

algoritmo DBSCAN utiliza uma abordagem baseada em densidade, na qual a alteração dos centroides não necessariamente afetaria os resultados de agrupamento. Dessa forma, o único fator aleatório dentro da análise de estabilidade dos dados da base de dados de SNPs é a escolha do subconjunto de amostras instáveis que preserva a estrutura dos dados por parte do algoritmo genético. Considerando estas questões, foi realizado apenas um teste com a remoção de todas as amostras e foram realizados cinco testes de seleção e remoção com o algoritmo genético.

A Tabela 4.4 apresenta a distribuição das amostras instáveis e das amostras que isoladamente preservam a estrutura dos dados entre os grupos da base de dados de SNPs. A partir dos resultados da análise de estabilidade dos dados, é possível notar que as amostras das raças Nelore e Jersey são consideradas estáveis, uma vez que não existem amostras consideradas como instáveis entre as amostras da raça Nelore e apenas três amostras (0,29%) da raça Jersey são consideradas instáveis. Por outro lado, praticamente todas as amostras da raça Holandesa (99,65%) são consideradas instáveis, sendo que aproximadamente 84% das amostras da raça Holandesa isoladamente preservam a estrutura dos dados e, portanto, seriam removidas segundo a abordagem de Mulder (2014).

Tabela 4.4: Distribuição das amostras instáveis e que isoladamente preservam a estrutura dos dados entre os grupos da base de dados de SNPs. Os valores correspondentes à raça Holandesa indicam que esta é a raça mais instável e praticamente 84,06% de suas amostras preservam a estrutura dos dados. Por outro lado, as raças Nelore e Jersey são consideradas estáveis.

Grupo	Instáveis	Preservam estrutura
Holandesa	**575 (99,65%)**	**485 (84,06%)**
Jersey	3 (0,29%)	2 (0,20%)
Nelore	0	0

As Tabelas 4.5 (Figura 4.8) e 4.6 (Figura 4.9) apresentam a instabilidade resultante após a remoção das amostras instáveis que preservam a estrutura da base de dados de SNPs utilizando os seguintes critérios:

i) identificação e remoção de todas as amostras instáveis (Equação 2.14, pg. 31) que isoladamente preservam a estrutura dos dados (Equação 2.16, pg. 32), proposto por Mulder (2014);

ii) seleção e remoção proposta neste trabalho, no qual um subconjunto de amostras instáveis (Equação 2.14, pg. 31) que preserva a estrutura dos dados (Equação 3.1, pg. 48) é selecionado e removido utilizando algoritmo genético com duas funções de avaliação.

A Tabela 4.5 e a Figura 4.8 apresentam os resultados obtidos utilizando a primeira função de avaliação, $f(x)$, para o algoritmo genético (Equação 3.3, pg. 49) e a Tabela 4.6 e a Figura 4.9 apresentam os resultados obtidos utilizando a segunda função de avaliação, $g(x)$, (Equação 3.4, pg. 49) para o algoritmo genético.

Tabela 4.5: Instabilidade resultante após a remoção de amostras instáveis que preservam a estrutura da base de dados de SNPs utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a primeira função de avaliação (Equação 3.3). A remoção de todas as amostras apresentou resultado de instabilidade ($19{,}81 \times 10^{-4}$) pior que a instabilidade original ($7{,}28 \times 10^{-4}$) da base de dados de SNPs. Por outro lado, a seleção e remoção utilizando algoritmo genético apresentou melhores resultados de instabilidade em todos os testes.

Teste id	Todas	AG $f(x)$	Todas − AG $f(x)$
1	$19{,}81 \times 10^{-4}$	$\mathbf{1{,}36 \times 10^{-4}}$	$18{,}45 \times 10^{-4}$
2		$\mathbf{1{,}34 \times 10^{-4}}$	$18{,}47 \times 10^{-4}$
3		$\mathbf{1{,}36 \times 10^{-4}}$	$18{,}45 \times 10^{-4}$
4		$\mathbf{1{,}35 \times 10^{-4}}$	$18{,}46 \times 10^{-4}$
5		$\mathbf{1{,}38 \times 10^{-4}}$	$18{,}43 \times 10^{-4}$

Considerando a instabilidade resultante após a seleção e remoção de amostras instáveis que preservam a estrutura dos dados utilizando a primeira função de avaliação (Tabela 4.5 e Figura 4.8), é possível notar que em todos os testes realizados a abordagem de seleção e remoção utilizando algoritmo genético apresenta resultados de instabilidade melhores que a instabilidade original ($\mu_D(A) = 7{,}28 \times 10^{-4}$). A diferença entre tais resultados (AG $f(x)$) e a instabilidade original ($\mu_D(A)$) está no intervalo $5{,}90 \times 10^{-4}$ (teste 5) a $5{,}94 \times 10^{-4}$ (teste 2). Por outro lado, a remoção de todas as amostras apresenta considerável aumento de instabilidade ($7{,}28 \times 10^{-4}$ para $19{,}81 \times 10^{-4}$), isto é, aumento de $12{,}53 \times 10^{-4}$.

Considerando a instabilidade resultante após a seleção e remoção de amostras instáveis que preservam a estrutura dos dados utilizando a segunda função de avaliação (Tabela 4.6 e Figura 4.9), é possível notar que em todos os testes realizados a abordagem de seleção e

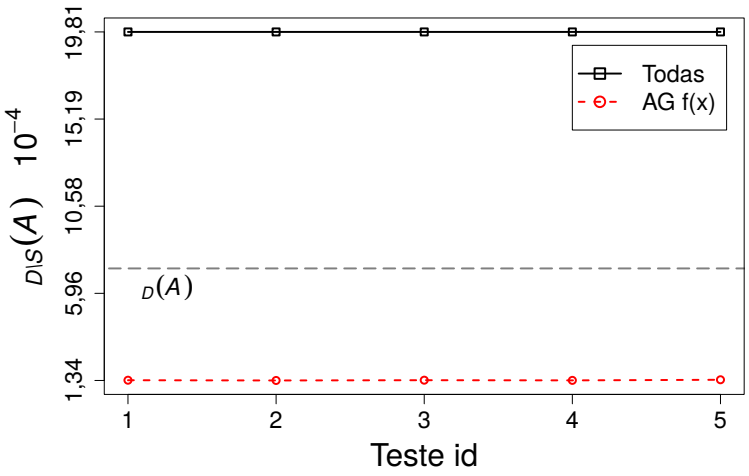

Figura 4.8: Instabilidade resultante após a remoção de amostras instáveis que preservam a estrutura da base de dados de SNPs utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a primeira função de avaliação (Equação 3.3). A remoção de todas as amostras apresentou resultado de instabilidade ($19{,}81 \times 10^{-4}$) pior que a instabilidade original ($7{,}28 \times 10^{-4}$) da base de dados de SNPs. Por outro lado, a seleção e remoção utilizando algoritmo genético apresentou melhores resultados de instabilidade em todos os testes. A linha tracejada representa a instabilidade ($\mu_D(A)$) da base de dados sem remoção de amostras.

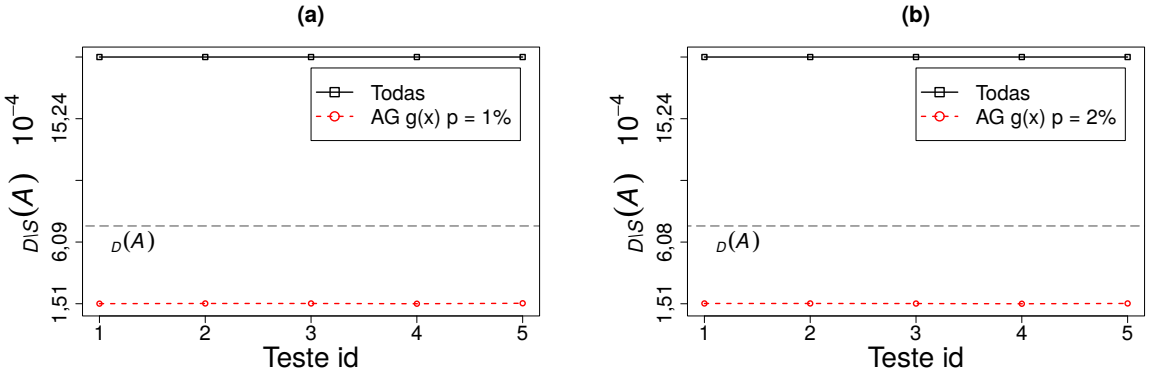

Figura 4.9: Instabilidade resultante após a remoção de amostras instáveis que preservam a estrutura da base de dados de SNPs utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a segunda função de avaliação (Equação 3.4) com $p = 1\%$ (a) e $p = 2\%$ (b). A remoção de todas as amostras apresentou resultado de instabilidade ($19{,}81 \times 10^{-4}$) pior que a instabilidade original ($7{,}28 \times 10^{-4}$) da base de dados de SNPs. Por outro lado, a seleção e remoção utilizando algoritmo genético apresentou melhores resultados de instabilidade em todos os testes. A linha tracejada representa a instabilidade ($\mu_D(A)$) da base de dados sem remoção de amostras.

Tabela 4.6: Instabilidade resultante após a remoção de amostras instáveis que preservam a estrutura da base de dados de SNPs utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a segunda função de avaliação (Equação 3.4). A remoção de todas as amostras apresentou resultado de instabilidade ($19{,}81 \times 10^{-4}$) pior que a instabilidade original ($7{,}28 \times 10^{-4}$) da base de dados de SNPs. Por outro lado, a seleção e remoção utilizando algoritmo genético apresentou melhores resultados de instabilidade em todos os testes.

p	k	c	Teste id	Todas	AG $g(x)$	Todas − AG $g(x)$
1%	2	0,201	1	$19{,}81 \times 10^{-4}$	$\mathbf{1{,}52 \times 10^{-4}}$	$18{,}29 \times 10^{-4}$
			2		$\mathbf{1{,}53 \times 10^{-4}}$	$18{,}28 \times 10^{-4}$
			3		$\mathbf{1{,}53 \times 10^{-4}}$	$18{,}28 \times 10^{-4}$
			4		$\mathbf{1{,}51 \times 10^{-4}}$	$18{,}30 \times 10^{-4}$
			5		$\mathbf{1{,}54 \times 10^{-4}}$	$18{,}27 \times 10^{-4}$
2%	2	0,099	1	$19{,}81 \times 10^{-4}$	$\mathbf{1{,}53 \times 10^{-4}}$	$18{,}28 \times 10^{-4}$
			2		$\mathbf{1{,}53 \times 10^{-4}}$	$18{,}28 \times 10^{-4}$
			3		$\mathbf{1{,}52 \times 10^{-4}}$	$18{,}29 \times 10^{-4}$
			4		$\mathbf{1{,}51 \times 10^{-4}}$	$18{,}30 \times 10^{-4}$
			5		$\mathbf{1{,}53 \times 10^{-4}}$	$18{,}28 \times 10^{-4}$

remoção utilizando algoritmo genético ainda apresenta resultados de instabilidade melhores que a instabilidade original ($\mu_D(A) = 7{,}28 \times 10^{-4}$). Nesse caso, a diferença entre tais resultados (AG $g(x)$) e a instabilidade original ($\mu_D(A)$), com $p = 1\%$ está no intervalo $5{,}74 \times 10^{-4}$ (teste 5) a $5{,}77 \times 10^{-4}$ (teste 4) e com $p = 2\%$ está no intervalo $5{,}75 \times 10^{-4}$ (testes 1, 2 e 5) a $5{,}77 \times 10^{-4}$ (teste 4).

Uma vez que se conhece a instabilidade resultante após a remoção de amostras instáveis que preservam a estrutura dos dados utilizando ambas as abordagens, a próxima análise avalia a distribuição das amostras removidas entre os grupos com o propósito de verificar se foi possível reduzir o número de amostras removidas e quanto essa redução afetou a instabilidade resultante.

4.1.6 DISTRIBUIÇÃO DAS AMOSTRAS REMOVIDAS

As Tabelas 4.7 , 4.8 e 4.9 apresentam a distribuição das amostras instáveis que preservam a estrutura dos dados removidas entre os grupos da base de dados de SNPs utilizando os seguintes critérios:

i) identificação e remoção de todas as amostras instáveis (Equação 2.14, pg. 31) que isoladamente preservam a estrutura dos dados (Equação 2.16, pg. 32), proposto por Mulder (2014);

ii) seleção e remoção proposta neste trabalho, no qual um subconjunto de amostras instáveis (Equação 2.14, pg. 31) que preserva a estrutura dos dados (Equação 3.1, pg. 48) é selecionado e removido utilizando algoritmo genético com duas funções de avaliação.

A Tabela 4.7 apresenta os resultados obtidos utilizando a primeira função de avaliação proposta para o algoritmo genético (Equação 3.3, pg. 49).

Tabela 4.7: Distribuição das amostras instáveis que preservam a estrutura dos dados selecionadas e removidas entre os grupos da base de dados de SNPs utilizando a primeira função de avaliação (Equação 3.3). A seleção e remoção de amostras utilizando algoritmo genético apresentou considerável redução no percentual de amostras removidas em todos os testes.

Grupo	Teste id	Todas	AG $f(x)$	Todas − AG $f(x)$
Holandesa	1	485 (84,06%)	**166 (28,77%)**	319 (55,29%)
Jersey		2 (0,20%)	0	2 (0,20%)
Nelore		0	0	0
Holandesa	2	485 (84,06%)	**174 (30,16%)**	311 (53,90%)
Jersey		2 (0,20%)	0	2 (0,20%)
Nelore		0	0	0
Holandesa	3	485 (84,06%)	**167 (28,94%)**	318 (55,12%)
Jersey		2 (0,20%)	0	2 (0,20%)
Nelore		0	0	0
Holandesa	4	485 (84,06%)	**168 (29,12%)**	317 (54,94%)
Jersey		2 (0,20%)	1 (0,10%)	1 (0,10%)
Nelore		0	0	0
Holandesa	5	485 (84,06%)	**157 (27,21%)**	328 (56,85%)
Jersey		2 (0,20%)	0	2 (0,20%)
Nelore		0	0	0

Considerando a distribuição das amostras instáveis que preservam a estrutura dos dados selecionadas e removidas utilizando a primeira função de avaliação (Tabela 4.7), é possível notar que utilizando algoritmo genético em todos os testes foi possível reduzir

consideravelmente o número de amostras removidas em relação à estratégia de remoção de todas as amostras. Utilizando o algoritmo genético com a primeira função de avaliação, foram alcançadas reduções no percentual de amostras removidas da raça Holandesa entre 53,9% (teste 2) e 56,85% (teste 5).

Analisando a instabilidade resultante após a seleção e remoção de amostras utilizando a primeira função de avaliação (Tabela 4.5 e Figura 4.8) em relação à distribuição dessas amostras (Tabela 4.7), nota-se que com reduções entre $5{,}90 \times 10^{-4}$ (teste 5) e $5{,}94 \times 10^{-4}$ (teste 2) na instabilidade original ($\mu_D(A) = 7{,}28 \times 10^{-4}$), a abordagem proposta utilizando algoritmo genético conseguiu reduções de 53,9% a 56,85% no percentual de amostras removidas da raça Holandesa.

A Tabela 4.8 apresenta os resultados obtidos utilizando a segunda função de avaliação (Equação 3.4, pg. 49) com o percentual de redução p igual a 1% e o coeficiente k igual a 2.

Tabela 4.8: Distribuição das amostras instáveis que preservam a estrutura dos dados selecionadas e removidas entre os grupos da base de dados de SNPs utilizando a segunda função de avaliação com $p = 1\%$ e $k = 2$ (Equação 3.4). A seleção e remoção de amostras utilizando algoritmo genético apresentou considerável redução no percentual de amostras removidas em todos os testes.

Grupo	Teste id	Todas	AG $g(x)$ $p = 1\%$	Todas − AG $g(x)$
Holandesa	1	485 (84,06%)	**87 (15,08%)**	398 (68,26%)
Jersey		2 (0,20%)	0	2 (0,20%)
Nelore		0	0	0
Holandesa	2	485 (84,06%)	**82 (14,21%)**	403 (69,85%)
Jersey		2 (0,20%)	0	2 (0,20%)
Nelore		0	0	0
Holandesa	3	485 (84,06%)	**81 (14,04%)**	404 (70,02%)
Jersey		2 (0,20%)	0	2 (0,20%)
Nelore		0	0	0
Holandesa	4	485 (84,06%)	**92 (15,94%)**	393 (68,12%)
Jersey		2 (0,20%)	0	2 (0,20%)
Nelore		0	0	0
Holandesa	5	485 (84,06%)	**76 (13,17%)**	409 (70,89%)
Jersey		2 (0,20%)	0	2 (0,20%)
Nelore		0	0	0

Considerando a distribuição das amostras instáveis que preservam a estrutura dos dados selecionadas e removidas utilizando a segunda função de avaliação e $p = 1\%$ (Tabela 4.8), é possível notar que utilizando algoritmo genético em todos os testes ainda foi possível reduzir consideravelmente o número de amostras removidas em relação à estratégia de remoção de todas as amostras. Utilizando o algoritmo genético com a segunda função de avaliação e $p = 1\%$, foram alcançadas reduções no percentual de amostras removidas da raça Holandesa entre 68,12% (teste 4) e 70,89% (teste 5).

Analisando a instabilidade resultante após a seleção e remoção de amostras utilizando a segunda função de avaliação e $p = 1\%$ (Tabela 4.6 e Figura 4.9) em relação à distribuição dessas amostras (Tabela 4.8), nota-se que com reduções entre $5{,}74 \times 10^{-4}$ (teste 5) e $5{,}77 \times 10^{-4}$ (teste 4) na instabilidade original ($\mu_D(A) = 7{,}28 \times 10^{-4}$), a abordagem proposta utilizando algoritmo genético conseguiu reduções de 68,12% a 70,89% no percentual de amostras removidas da raça Holandesa. Isto é, utilizando a segunda função de avaliação e definindo que o número de amostras para remoção n só poderia ser multiplicado por 2 ($k = 2$) caso houvesse redução de pelo menos 1% ($p = 1\%$) na instabilidade resultante, foi possível alcançar reduções de 68,12% a 70,89% no percentual de amostras removidas da raça Holandesa.

A Tabela 4.9 apresenta os resultados obtidos utilizando ainda a segunda função de avaliação, com o percentual de redução p igual a 2% e o coeficiente k igual a 2.

Considerando a distribuição das amostras instáveis que preservam a estrutura dos dados selecionadas e removidas utilizando a segunda função de avaliação e $p = 2\%$ (Tabela 4.9), é possível notar que utilizando algoritmo genético em todos os testes ainda foi possível reduzir consideravelmente o número de amostras removidas em relação à estratégia de remoção de todas as amostras. Utilizando o algoritmo genético com a segunda função de avaliação e $p = 2\%$, foram alcançadas reduções no percentual de amostras removidas da raça Holandesa entre 68,46% (teste 4) e 70,02% (teste 1).

Analisando a instabilidade resultante após a seleção e remoção de amostras utilizando a segunda função de avaliação e $p = 2\%$ (Tabela 4.6 e Figura 4.9) em relação à distribuição dessas amostras (Tabela 4.9), nota-se que com reduções entre $5{,}75 \times 10^{-4}$ (testes 1, 2 e 5) e $5{,}77 \times 10^{-4}$ (teste 4) na instabilidade original ($\mu_D(A) = 7{,}28 \times 10^{-4}$), a abordagem proposta com algoritmo genético conseguiu reduções de 68,46% a 70,02% no percentual de amostras removidas da raça Holandesa. Isto é, utilizando a segunda função de avaliação

Tabela 4.9: Distribuição das amostras instáveis que preservam a estrutura dos dados selecionadas e removidas entre os grupos da base de dados de SNPs utilizando a segunda função de avaliação com $p = 2\%$ e $k = 2$ (Equação 3.4). A seleção e remoção de amostras utilizando algoritmo genético apresentou considerável redução no percentual de amostras removidas em todos os testes.

Grupo	Teste id	Todas	AG $g(x)$ $p = 2\%$	Todas − AG $g(x)$
Holandesa	1	485 (84,06%)	**81 (14,04%)**	404 (70,02%)
Jersey		2 (0,20%)	0	2 (0,20%)
Nelore		0	0	0
Holandesa	2	485 (84,06%)	**82 (14,21%)**	403 (69,85%)
Jersey		2 (0,20%)	0	2 (0,20%)
Nelore		0	0	0
Holandesa	3	485 (84,06%)	**85 (14,73%)**	400 (69,33%)
Jersey		2 (0,20%)	0	2 (0,20%)
Nelore		0	0	0
Holandesa	4	485 (84,06%)	**90 (15,60%)**	395 (68,46%)
Jersey		2 (0,20%)	0	2 (0,20%)
Nelore		0	0	0
Holandesa	5	485 (84,06%)	**83 (14,38%)**	402 (69,68%)
Jersey		2 (0,20%)	0	2 (0,20%)
Nelore		0	0	0

e definindo que o número de amostras para remoção n só poderia ser multiplicado por 2 ($k = 2$) caso houvesse redução de pelo menos 2% ($p = 2\%$) na instabilidade resultante, foi possível alcançar reduções de 68,46% a 70,02% no percentual de amostras removidas da raça Holandesa.

4.2 IRIS

Uma vez que a base de dados *Iris* foi utilizada apenas com o intuito de comparar a abordagem aqui proposta com a abordagem apresentada por Mulder (2014) e devido ao número de instâncias analisadas (Seção 3.3.4), os resultados de agrupamento não são apresentados em detalhes, sendo apresentados aqui os resultados da análise de estabilidade do agrupamento e da análise de estabilidade dos dados.

4.2.1 CSV E INSTABILIDADE

A partir dos resultados de agrupamento da base de dados *Iris* apresentados pelo algoritmo K-*means* foi possível realizar a análise de estabilidade do agrupamento. Os resultados dessa análise indicaram $\mu_D(A) = 1{,}24 \times 10^{-1}$ para a instabilidade e $CSV_D(A) = 0{,}81 \times 10^{-1}$ para a variância de estabilidade do agrupamento. Após a análise de estabilidade de agrupamento foram realizadas análises de estabilidade dos dados.

4.2.2 SELEÇÃO E REMOÇÃO DE AMOSTRAS PREJUDICIAIS

Como citado no exemplo ilustrativo apresentado na Seção 2.3.5, as amostras instáveis que preservam a estrutura dos dados são identificadas por meio de uma nova aplicação do algoritmo de agrupamento sobre a base de dados desconsiderando-se tais amostras. Como citado na Seção 3.3.4, o algoritmo K-*means* foi aplicado sobre essa base de dados utilizando inicialização aleatória para os centroides. Considerando estas questões, cada aplicação do algoritmo K-*means* sobre a base de dados para identificar as amostras que preservam a estrutura dos dados pode apresentar diferentes resultados, sendo assim, foram realizados cinco testes para identificar tais amostras.

A Tabela 4.10 apresenta a distribuição das amostras instáveis e das amostras que isoladamente preservam a estrutura dos dados entre os grupos da base de dados *Iris*. A partir dos resultados da análise de estabilidade dos dados, é possível notar que não existem amostras consideradas como instáveis entre as amostras do grupo *Iris setosa*. Por outro lado, todas as amostras dos grupos *Iris versicolor* e *Iris virginica* são consideradas instáveis, sendo que aproximadamente 25% das amostras de cada grupo, isoladamente preservam a estrutura dos dados. Sendo assim, as amostras do grupo *Iris setosa* podem ser consideradas estáveis, o que não ocorre para as amostras dos demais grupos, e no pior caso, dois grupos teriam aproximadamente 25% de suas amostras removidas por serem consideradas instáveis e isoladamente preservarem a estrutura dos dados, segundo a abordagem de Mulder (2014).

As Tabelas 4.11 (Figura 4.10) e 4.12 (Figura 4.11) apresentam a instabilidade resultante após a remoção das amostras instáveis que preservam a estrutura da base de dados *Iris* utilizando os seguintes critérios:

i) identificação e remoção de todas as amostras instáveis (Equação 2.14, pg. 31) que

Tabela 4.10: Distribuição das amostras instáveis e que isoladamente preservam a estrutura dos dados entre os grupos da base de dados *Iris*. Os valores correspondentes ao grupo setosa indicam que todas as amostras desse grupo são estáveis. No entanto, os valores correspondentes ao grupos versicolor e virginica indicam que todas as amostras desses grupos são instáveis e aproximadamente 25% das amostras de cada grupo, isoladamente preservam a estrutura dos dados.

Grupo	Teste id	Instáveis	Preservam estrutura
setosa		**0**	**0**
versicolor	1	50 (100%)	13 (26%)
virginica		50 (100%)	12 (24%)
setosa		**0**	**0**
versicolor	2	50 (100%)	11 (22%)
virginica		50 (100%)	10 (20%)
setosa		**0**	**0**
versicolor	3	50 (100%)	15 (30%)
virginica		50 (100%)	13 (26%)
setosa		**0**	**0**
versicolor	4	50 (100%)	11 (22%)
virginica		50 (100%)	12 (24%)
setosa		**0**	**0**
versicolor	5	50 (100%)	15 (30%)
virginica		50 (100%)	14 (28%)

isoladamente preservam a estrutura dos dados (Equação 2.17, pg. 32), proposto por Mulder (2014);

ii) seleção e remoção proposta neste trabalho, no qual um subconjunto de amostras instáveis (Equação 2.14, pg. 31) que preserva a estrutura dos dados (Equação 3.2, pg. 48) é selecionado e removido utilizando algoritmo genético com duas funções de avaliação.

A Tabela 4.11 e a Figura 4.10 apresentam os resultados obtidos utilizando a primeira função de avaliação, $f(x)$, para o algoritmo genético (Equação 3.3, pg. 49) e a Tabela 4.12 e a Figura 4.11 apresentam os resultados obtidos utilizando a segunda função de avaliação, $g(x)$, (Equação 3.4, pg. 49) para o algoritmo genético.

Considerando a instabilidade resultante após a seleção e remoção de amostras instáveis

Tabela 4.11: Instabilidade resultante após a seleção e remoção de amostras instáveis que preservam a estrutura da base de dados *Iris* utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a primeira função de avaliação (Equação 3.3). A remoção de todas as amostras apresentou melhores resultados de instabilidade em um teste (teste 1) enquanto a seleção e remoção utilizando algoritmo genético apresentou melhores resultados em quatro testes.

Teste id	Todas	AG $f(x)$	Todas − AG $f(x)$
1	$\mathbf{0{,}95 \times 10^{-1}}$	$0{,}96 \times 10^{-1}$	$-0{,}01 \times 10^{-1}$
2	$1{,}21 \times 10^{-1}$	$\mathbf{0{,}95 \times 10^{-1}}$	$0{,}26 \times 10^{-1}$
3	$1{,}06 \times 10^{-1}$	$\mathbf{0{,}96 \times 10^{-1}}$	$0{,}10 \times 10^{-1}$
4	$1{,}12 \times 10^{-1}$	$\mathbf{1{,}00 \times 10^{-1}}$	$0{,}12 \times 10^{-1}$
5	$1{,}14 \times 10^{-1}$	$\mathbf{0{,}92 \times 10^{-1}}$	$0{,}22 \times 10^{-1}$

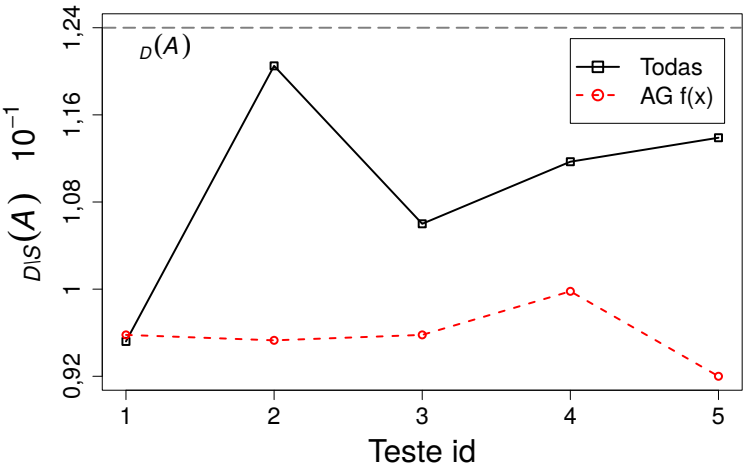

Figura 4.10: Instabilidade resultante após a seleção e remoção de amostras instáveis que preservam a estrutura da base de dados *Iris* utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a primeira função de avaliação (Equação 3.3). A remoção de todas as amostras apresentou melhores resultados de instabilidade em um teste (teste 1) enquanto a seleção e remoção utilizando algoritmo genético apresentou melhores resultados em quatro testes. A linha tracejada representa a instabilidade ($\mu_D(A)$) da base de dados sem remoção de amostras.

que preservam a estrutura dos dados utilizando a primeira função de avaliação (Tabela 4.11 e Figura 4.10), é possível notar que em quatro (80%) dos testes realizados (testes 2, 3, 4 e 5) a abordagem de seleção e remoção utilizando algoritmo genético apresenta resultados de instabilidade melhores que os resultados apresentados pela remoção de todas as amostras.

Sendo que a diferença entre os resultados de instabilidade apresentados pela remoção de todas as amostras e pela seleção e remoção utilizando algoritmo genético está no intervalo $-0{,}01 \times 10^{-1}$ (teste 1) a $0{,}26 \times 10^{-1}$ (teste 2).

Tabela 4.12: Instabilidade resultante após a seleção e remoção de amostras instáveis que preservam a estrutura da base de dados *Iris* utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a segunda função de avaliação (Equação 3.4). A remoção de todas a amostras apresentou melhores resultados de instabilidade em um teste (teste 1) no qual a segunda função de avaliação utilizou $p = 2\%$. Nos demais testes a seleção e remoção utilizando algoritmo genético apresentou melhores resultados.

p	k	c	Teste id	Todas	AG $g(x)$	Todas − AG $g(x)$
1%	2	3,920	1	$0{,}95 \times 10^{-1}$	$\mathbf{0{,}93 \times 10^{-1}}$	$0{,}02 \times 10^{-1}$
		4,667	2	$1{,}21 \times 10^{-1}$	$\mathbf{1{,}03 \times 10^{-1}}$	$0{,}18 \times 10^{-1}$
		3,500	3	$1{,}06 \times 10^{-1}$	$\mathbf{0{,}96 \times 10^{-1}}$	$0{,}10 \times 10^{-1}$
		4,261	4	$1{,}12 \times 10^{-1}$	$\mathbf{0{,}98 \times 10^{-1}}$	$0{,}14 \times 10^{-1}$
		3,379	5	$1{,}14 \times 10^{-1}$	$\mathbf{0{,}95 \times 10^{-1}}$	$0{,}19 \times 10^{-1}$
2%	2	1,920	1	$\mathbf{0{,}95 \times 10^{-1}}$	$1{,}05 \times 10^{-1}$	$-0{,}10 \times 10^{-1}$
		2,286	2	$1{,}21 \times 10^{-1}$	$\mathbf{1{,}04 \times 10^{-1}}$	$0{,}17 \times 10^{-1}$
		1,714	3	$1{,}06 \times 10^{-1}$	$\mathbf{1{,}01 \times 10^{-1}}$	$0{,}05 \times 10^{-1}$
		2,087	4	$1{,}12 \times 10^{-1}$	$\mathbf{1{,}03 \times 10^{-1}}$	$0{,}09 \times 10^{-1}$
		1,655	5	$1{,}14 \times 10^{-1}$	$\mathbf{0{,}96 \times 10^{-1}}$	$0{,}18 \times 10^{-1}$

Considerando a instabilidade resultante após a seleção e remoção de amostras instáveis que preservam a estrutura dos dados utilizando a segunda função de avaliação (Tabela 4.12 e Figura 4.11), é possível notar que a abordagem de seleção e remoção utilizando algoritmo genético apresenta resultados de instabilidade melhores que os resultados apresentados pela remoção de todas as amostras na maioria do testes, em um teste (teste 1, $p = 2\%$) a remoção de todas as amostras apresenta resultado melhor. Sendo que a diferença entre os resultados de instabilidade apresentados pela remoção de todas as amostras e pela seleção e remoção utilizando algoritmo genético, com *p = 1%* está no intervalo $0{,}02 \times 10^{-1}$ (teste 1) a $0{,}19 \times 10^{-1}$ (teste 5) e com *p = 2%* está no intervalo de $-0{,}10 \times 10^{-1}$ (teste 1) a $0{,}18 \times 10^{-1}$ (teste 5).

Uma vez que se conhece a instabilidade resultante após a remoção de amostras instáveis que preservam a estrutura dos dados utilizando ambas as abordagens, a próxima análise avalia a distribuição das amostras removidas entre os grupos com o propósito de verificar

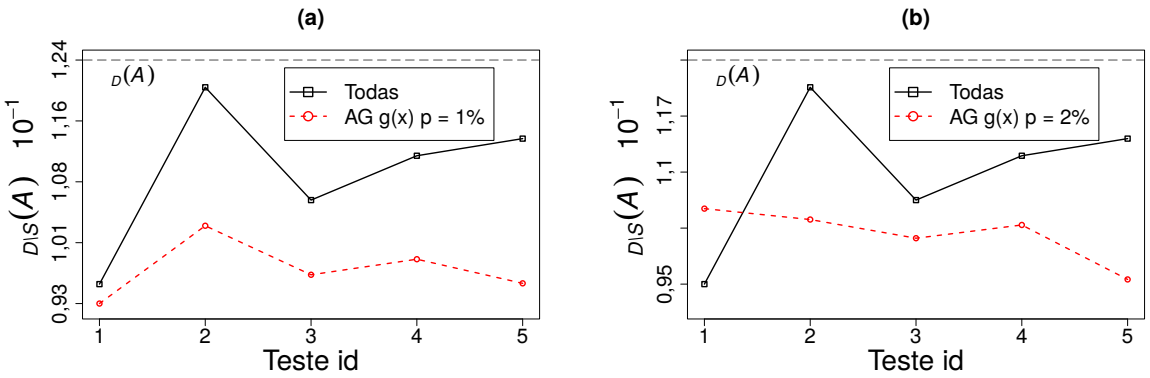

Figura 4.11: Instabilidade resultante após a seleção e remoção de amostras instáveis que preservam a estrutura da base de dados *Iris* utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a segunda função de avaliação (Equação 3.4) com $p = 1\%$ (a) e $p = 2\%$ (b). A remoção de todas a amostras apresentou melhores resultados de instabilidade em um teste (teste 1) no qual a segunda função de avaliação utilizou $p = 2\%$. Nos demais testes a seleção e remoção utilizando algoritmo genético apresentou melhores resultados. A linha tracejada representa a instabilidade ($\mu_D(A)$) da base de dados sem remoção de amostras.

se foi possível reduzir o número de amostras removidas e quanto essa redução afetou a instabilidade resultante.

4.2.3 DISTRIBUIÇÃO DAS AMOSTRAS REMOVIDAS

As Tabelas 4.13 , 4.14 e 4.15 apresentam a distribuição das amostras instáveis que preservam a estrutura dos dados removidas entre os grupos da base de dados *Iris* utilizando os seguintes critérios:

i) identificação e remoção de todas as amostras instáveis (Equação 2.14, pg. 31) que isoladamente preservam a estrutura dos dados (Equação 2.17, pg. 32), proposto por Mulder (2014);

ii) seleção e remoção proposta neste trabalho, no qual um subconjunto de amostras instáveis (Equação 2.14, pg. 31) que preserva a estrutura dos dados (Equação 3.2, pg. 48) é selecionado e removido utilizando algoritmo genético com duas funções de avaliação.

A Tabela 4.13 apresenta os resultados obtidos utilizando a primeira função de avaliação proposta para o algoritmo genético (Equação 3.3, pg. 49).

Tabela 4.13: Distribuição das amostras instáveis que preservam a estrutura dos dados selecionadas e removidas entre os grupos da base de dados *Iris* utilizando a primeira função de avaliação (Equação 3.3). A seleção e remoção de amostras utilizando algoritmo genético apresentou redução no percentual de amostras removidas em todos os testes.

Grupo	Teste id	Todas	AG $f(x)$	Todas − AG $f(x)$
setosa	1	0	0	0
versicolor		13 (26%)	**9 (18%)**	4 (8%)
virginica		12 (24%)	**10 (20%)**	2 (4%)
setosa	2	0	0	0
versicolor		11 (22%)	**9 (18%)**	2 (4%)
virginica		10 (20%)	**7 (14%)**	3 (6%)
setosa	3	0	0	0
versicolor		15 (30%)	**9 (18%)**	6 (12%)
virginica		13 (26%)	**10 (20%)**	3 (6%)
setosa	4	0	0	0
versicolor		11 (22%)	**6 (12%)**	5 (10%)
virginica		12 (24%)	**8 (16%)**	4 (8%)
setosa	5	0	0	0
versicolor		15 (30%)	**9 (18%)**	6 (12%)
virginica		14 (28%)	**11 (22%)**	3 (6%)

Considerando a distribuição das amostras instáveis que preservam a estrutura dos dados selecionadas e removidas utilizando a primeira função de avaliação (Tabela 4.13), é possível notar que utilizando algoritmo genético em todos os testes foi possível reduzir o número de amostras removidas em relação à estratégia de remoção de todas as amostras. Quanto ao percentual de amostras removidas no grupo *Iris versicolor*, a redução alcançada está entre 4% (teste 2) e 12% (testes 3 e 5) e no grupo *Iris virginica* a redução alcançada está entre 4% (teste 1) e 8% (teste 4). No melhor caso, para o grupo *Iris versicolor* (testes 3 e 5), foi possível reduzir o percentual de 30% para 18% e para o grupo *Iris virginica* (teste 4) foi possível reduzir o percentual de 24% para 16%.

Analisando a instabilidade resultante após a seleção e remoção de amostras utilizando a primeira função de avaliação (Tabela 4.11 e Figura 4.10) em relação à distribuição dessas amostras (Tabela 4.13), nota-se que com reduções entre $0{,}10 \times 10^{-1}$ (teste 3) e $0{,}26 \times 10^{-1}$ (teste 2) – apesar do aumento de $0{,}01 \times 10^{-1}$ no teste 1 – na instabilidade apresentada pela

remoção de todas as amostras, a abordagem proposta com algoritmo genético conseguiu reduções de 4% a 12% no percentual de amostras removidas no grupo *Iris versicolor* e reduções de 4% a 8% no percentual de amostras removidas no grupo *Iris virginica*.

A Tabela 4.14 apresenta os resultados obtidos utilizando a segunda função de avaliação (Equação 3.4, pg. 49) com o percentual de redução p igual a 1% e o coeficiente k igual a 2.

Tabela 4.14: Distribuição das amostras instáveis que preservam a estrutura dos dados selecionadas e removidas entre os grupos da base de dados *Iris* utilizando a segunda função de avaliação com $p = 1\%$ e $k = 2$ (Equação 3.4). A seleção e remoção de amostras utilizando algoritmo genético apresentou redução no percentual de amostras removidas em todos os testes.

Grupo	Teste id	Todas	AG $g(x)$ $p = 1\%$	Todas − AG $g(x)$
setosa	1	0	0	0
versicolor		13 (26%)	**7 (14%)**	6 (12%)
virginica		12 (24%)	**8 (16%)**	4 (8%)
setosa	2	0	0	0
versicolor		11 (22%)	**5 (10%)**	6 (12%)
virginica		10 (20%)	**3 (6%)**	7 (14%)
setosa	3	0	0	0
versicolor		15 (30%)	**7 (14%)**	8 (16%)
virginica		13 (26%)	**8 (16%)**	5 (10%)
setosa	4	0	0	0
versicolor		11 (22%)	**7 (14%)**	4 (8%)
virginica		12 (24%)	**6 (12%)**	6 (12%)
setosa	5	0	0	0
versicolor		15 (30%)	**8 (16%)**	7 (14%)
virginica		14 (28%)	**8 (16%)**	6 (12%)

Considerando a distribuição das amostras instáveis que preservam a estrutura dos dados selecionadas e removidas utilizando a segunda função de avaliação e *p = 1%* (Tabela 4.14), é possível notar que utilizando algoritmo genético em todos os testes foi possível reduzir o número de amostras removidas em relação à estratégia de remoção de todas as amostras. Quanto ao percentual de amostras removidas no grupo *Iris versicolor*, a redução alcançada está entre 8% (teste 4) e 16% (teste 3) e no grupo *Iris virginica* a redução alcançada está entre 8% (teste 1) e 14% (teste 2). No melhor caso, para o grupo *Iris*

versicolor (teste 3), foi possível reduzir o percentual de remoção de amostras de 30% para 14% e para o grupo *Iris virginica* (teste 2) foi possível reduzir o percentual de 20% para 6%.

Analisando a instabilidade resultante após a seleção e remoção de amostras utilizando a segunda função de avaliação e $p = 1\%$ (Tabela 4.12 e Figura 4.11) em relação à distribuição dessas amostras (Tabela 4.14), nota-se que com reduções entre $0{,}02 \times 10^{-1}$ (teste 1) e $0{,}19 \times 10^{-1}$ (teste 5) na instabilidade apresentada pela remoção de todas as amostras, a abordagem proposta com algoritmo genético conseguiu reduções de 8% a 16% no percentual de amostras removidas no grupo *Iris versicolor* e reduções de 8% a 14% no percentual de amostras removidas no grupo *Iris virginica*. Isto é, utilizando a segunda função de avaliação e definindo que o número de amostras para remoção n só poderia ser multiplicado por 2 ($k = 2$) caso houvesse redução de pelo menos 1% ($p = 1\%$) na instabilidade resultante, foi possível alcançar reduções de 8% a 16% no percentual de amostras removidas no grupo *Iris versicolor* e reduções de 8% a 14% no percentual de amostras removidas no grupo *Iris virginica*.

A Tabela 4.15 apresenta os resultados obtidos utilizando ainda a segunda função de avaliação, com o percentual de redução p igual a 2% e o coeficiente k igual a 2.

Considerando a distribuição das amostras instáveis que preservam a estrutura dos dados selecionadas e removidas utilizando a segunda função de avaliação e $p = 2\%$ (Tabela 4.15), é possível notar que utilizando algoritmo genético em todos os testes foi possível reduzir consideravelmente o número de amostras removidas em relação à estratégia de remoção de todas as amostras. Quanto ao percentual de amostras removidas no grupo *Iris versicolor*, a redução alcançada está entre 16% (teste 2) e 22% (teste 1) e no grupo *Iris virginica* a redução alcançada está entre 12% (teste 2) e 18% (testes 1, 3 e 4). No melhor caso, para o grupo *Iris versicolor* (teste 1), foi possível reduzir o percentual de 26% para 4% e para o grupo *Iris virginica* (testes 1, 3 e 4) foi possível reduzir o percentual de 24% para 6%.

Analisando a instabilidade resultante após a seleção e remoção de amostras utilizando a segunda função de avaliação e $p = 2\%$ (Tabela 4.12 e Figura 4.11) em relação à distribuição dessas amostras (Tabela 4.15), nota-se que com reduções entre $0{,}05 \times 10^{-1}$ (teste 3) e $0{,}18 \times 10^{-1}$ (teste 5) – apesar do aumento de $0{,}1 \times 10^{-1}$ no teste 1 – na instabilidade apresentada pela remoção de todas as amostras, a abordagem proposta com algoritmo

Tabela 4.15: Distribuição das amostras instáveis que preservam a estrutura dos dados selecionadas e removidas entre os grupos da base de dados *Iris* utilizando a segunda função de avaliação com $p = 2\%$ e $k = 2$ (Equação 3.4). A seleção e remoção de amostras utilizando algoritmo genético apresentou considerável redução no percentual de amostras removidas em todos os testes.

Grupo	Teste id	Todas	AG $g(x)$ $p = 2\%$	Todas − AG $g(x)$
setosa		0	0	0
versicolor	1	13 (26%)	**2 (4%)**	11 (22%)
virginica		12 (24%)	**3 (6%)**	9 (18%)
setosa		0	0	0
versicolor	2	11 (22%)	**3 (6%)**	8 (16%)
virginica		10 (20%)	**4 (8%)**	6 (12%)
setosa		0	0	0
versicolor	3	15 (30%)	**6 (12%)**	9 (18%)
virginica		13 (26%)	**4 (8%)**	9 (18%)
setosa		0	0	0
versicolor	4	11 (22%)	**2 (4%)**	9 (18%)
virginica		12 (24%)	**3 (6%)**	9 (18%)
setosa		0	0	0
versicolor	5	15 (30%)	**6 (12%)**	9 (18%)
virginica		14 (28%)	**6 (12%)**	8 (16%)

genético conseguiu reduções de 16% a 22% no percentual de amostras removidas no grupo *Iris versicolor* e reduções de 12% a 18% no percentual de amostras removidas no grupo *Iris virginica*. Isto é, utilizando a segunda função de avaliação e definindo que o número de amostras para remoção n só poderia ser multiplicado por 2 ($k = 2$) caso houvesse redução de pelo menos 2% ($p = 2\%$) na instabilidade resultante, foi possível alcançar reduções de 16% a 22% no percentual de amostras removidas no grupo *Iris versicolor* e reduções de 12% a 18% no percentual de amostras removidas no grupo *Iris virginica*.

4.3 WINE

Assim como a base de dados *Iris*, a *Wine* foi utilizada apenas com o intuito de comparar a abordagem aqui proposta com a abordagem apresentada por Mulder (2014). Sendo assim, os resultados de agrupamento não são apresentados em detalhes, sendo apresentados aqui

os resultados da análise de estabilidade do agrupamento e da análise de estabilidade dos dados.

4.3.1 CSV E INSTABILIDADE

A partir dos resultados de agrupamento apresentados pelo algoritmo K-*means* foi possível realizar a análise de estabilidade do agrupamento. Os resultados dessa análise indicaram $\mu_D(A) = 1{,}38 \times 10^{-1}$ para a instabilidade e $CSV_D(A) = 0{,}91 \times 10^{-1}$ para a variância de estabilidade do agrupamento. Após a análise de estabilidade de agrupamento foram realizadas análises de estabilidade dos dados.

4.3.2 SELEÇÃO E REMOÇÃO DE AMOSTRAS PREJUDICIAIS

Como citado no exemplo ilustrativo apresentado na Seção 2.3.5, as amostras instáveis que preservam a estrutura dos dados são identificadas por meio de uma nova aplicação do algoritmo de agrupamento sobre a base de dados desconsiderando-se tais amostras. Como citado na Seção 3.3.4, o algoritmo K-*means* foi aplicado sobre essa base de dados utilizando inicialização aleatória para os centroides. Considerando estas questões, cada aplicação do algoritmo K-*means* sobre a base de dados para identificar as amostras que preservam a estrutura dos dados pode apresentar diferentes resultados, sendo assim, foram realizados cinco testes para identificar tais amostras.

A Tabela 4.16 apresenta a distribuição das amostras instáveis e das amostras que isoladamente preservam a estrutura dos dados entre os grupos da base de dados *Wine*. A partir dos resultados da análise de estabilidade dos dados, é possível notar que existem amostras instáveis que isoladamente preservam a estrutura dos dados em todos os grupos, sendo que o grupo *tipo 3* apresenta o percentual mais elevado. Enquanto os grupos *tipo 1* e *tipo 2* apresentam no pior caso, aproximadamente 35% de amostras instáveis, o grupo *tipo 3* apresenta 75% de amostras instáveis. Com respeito ao percentual de amostras que isoladamente preservam a estrutura dos dados, os grupos *tipo 1* e *tipo 2* apresentam, no pior caso, aproximadamente 34% de amostras que preservam a estrutura e o grupo *tipo 3* apresenta, no pior caso, aproximadamente 69% de amostras que preservam a estrutura dos dados. As amostras do grupo *tipo 1* podem ser consideradas aquelas que apresentam maior estabilidade enquanto as amostras do grupo *tipo 3* podem ser consideradas aquelas que apresentam maior instabilidade. No pior caso, o grupo *tipo 1* teria aproximadamente

15% de suas amostras removidas (teste 5), o grupo *tipo 2* teria aproximadamente 34% de suas amostras removidas (teste 5) e o grupo *tipo 3* teria aproximadamente 69% de suas amostras removidas (teste 5) por serem consideradas instáveis e isoladamente preservarem a estrutura dos dados, segundo a abordagem de Mulder (2014).

Tabela 4.16: Distribuição das amostras instáveis e que isoladamente preservam a estrutura dos dados entre os grupos da base de dados *Wine*. Os valores correspondentes ao grupo tipo 3 indicam que este é o grupo mais instável e até 68,75% de suas amostras preservam a estrutura dos dados.

Grupo	Teste id	Instáveis	Preservam estrutura
tipo 1	1	13 (22,03%)	7 (11,86%)
tipo 2		25 (35,21%)	19 (26,76%)
tipo 3		**36 (75%)**	**30 (62,5%)**
tipo 1	2	13 (22,03%)	6 (10,17%)
tipo 2		25 (35,21%)	18 (25,35%)
tipo 3		**36 (75%)**	**27 (56,25%)**
tipo 1	3	13 (22,03%)	5 (8,47%)
tipo 2		25 (35,21%)	18 (25,35%)
tipo 3		**36 (75%)**	**30 (62,5%)**
tipo 1	4	13 (22,03%)	8 (13,56%)
tipo 2		25 (35,21%)	19 (26,76%)
tipo 3		**36 (75%)**	**26 (54,17%)**
tipo 1	5	13 (22,03%)	9 (15,25%)
tipo 2		25 (35,21%)	24 (33,8%)
tipo 3		**36 (75%)**	**33 (68,75%)**

As Tabelas 4.17 (Figura 4.12) e 4.18 (Figura 4.13) apresentam a instabilidade resultante após a remoção das amostras instáveis que preservam a estrutura da base de dados *Wine* utilizando os seguintes critérios:

i) identificação e remoção de todas as amostras instáveis (Equação 2.14, pg. 31) que isoladamente preservam a estrutura dos dados (Equação 2.17, pg. 32), proposto por Mulder (2014);

ii) seleção e remoção proposta neste trabalho, no qual um subconjunto de amostras instáveis (Equação 2.14, pg. 31) que preserva a estrutura dos dados (Equação 3.2,

pg. 48) é selecionado e removido utilizando algoritmo genético com duas funções de avaliação.

A Tabela 4.17 e a Figura 4.12 apresentam os resultados obtidos utilizando a primeira função de avaliação, $f(x)$, para o algoritmo genético (Equação 3.3, pg. 49) e a Tabela 4.18 e a Figura 4.13 apresentam os resultados obtidos utilizando a segunda função de avaliação, $g(x)$, (Equação 3.4, pg. 49) para o algoritmo genético.

Tabela 4.17: Instabilidade resultante após a remoção de amostras instáveis que preservam a estrutura da base de dados *Wine* utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a primeira função de avaliação (Equação 3.3). A remoção de todas as amostras apresentou melhores resultados de instabilidade em dois testes (testes 3 e 5) enquanto a seleção e remoção utilizando algoritmo genético apresentou melhores resultados em três testes (testes 1, 2 e 4).

Teste id	Todas	AG $f(x)$	Todas − AG $f(x)$
1	$1{,}02 \times 10^{-1}$	$\mathbf{0{,}95 \times 10^{-1}}$	$0{,}07 \times 10^{-1}$
2	$1{,}18 \times 10^{-1}$	$\mathbf{0{,}97 \times 10^{-1}}$	$0{,}21 \times 10^{-1}$
3	$\mathbf{0{,}96 \times 10^{-1}}$	$1{,}03 \times 10^{-1}$	$-0{,}07 \times 10^{-1}$
4	$1{,}04 \times 10^{-1}$	$\mathbf{0{,}97 \times 10^{-1}}$	$0{,}07 \times 10^{-1}$
5	$\mathbf{0{,}91 \times 10^{-1}}$	$0{,}97 \times 10^{-1}$	$-0{,}06 \times 10^{-1}$

Considerando a instabilidade resultante após a seleção e remoção de amostras instáveis que preservam a estrutura dos dados utilizando a primeira função de avaliação (Tabela 4.17 e Figura 4.12), é possível notar que em 60% dos testes realizados (testes 1, 2 e 4) a abordagem de seleção e remoção utilizando algoritmo genético apresenta resultados de instabilidade melhores que os resultados apresentados pela remoção de todas as amostras. Sendo que a diferença entre os resultados de instabilidade apresentados pela remoção de todas as amostras e pela seleção e remoção utilizando algoritmo genético está no intervalo de $-0{,}07 \times 10^{-1}$ (teste 3) a $0{,}21 \times 10^{-1}$ (teste 2).

Considerando a instabilidade resultante após a seleção e remoção de amostras instáveis que preservam a estrutura dos dados utilizando a segunda função de avaliação (Tabela 4.18 e Figura 4.13), é possível notar que em 40% dos testes realizados (testes 2 e 4, para ambos os valores de p) a abordagem de seleção e remoção utilizando algoritmo genético apresenta resultados de instabilidades melhores que os resultados apresentados pela remoção de todas as amostras. Sendo que a diferença entre os resultados de instabilidade apresentados

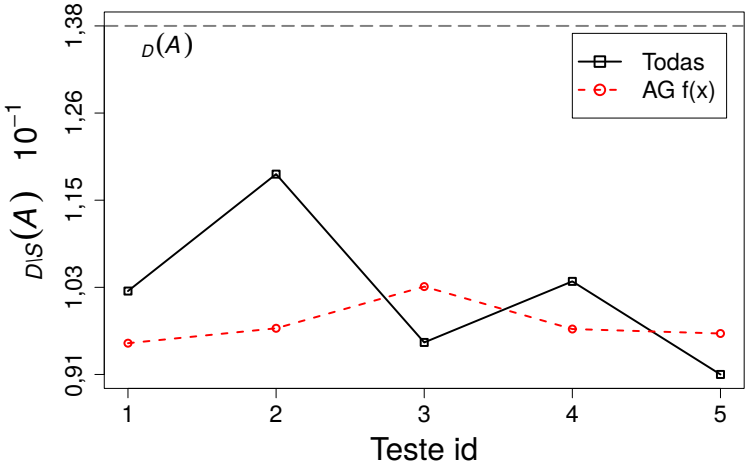

Figura 4.12: Instabilidade resultante após a remoção de amostras instáveis que preservam a estrutura da base de dados *Wine* utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a primeira função de avaliação (Equação 3.3). A remoção de todas as amostras apresentou melhores resultados de instabilidade em dois testes (testes 3 e 5) enquanto a seleção e remoção utilizando algoritmo genético apresentou melhores resultados em três testes (testes 1, 2 e 4). A linha tracejada representa a instabilidade ($\mu_D(A)$) da base de dados sem remoção de amostras.

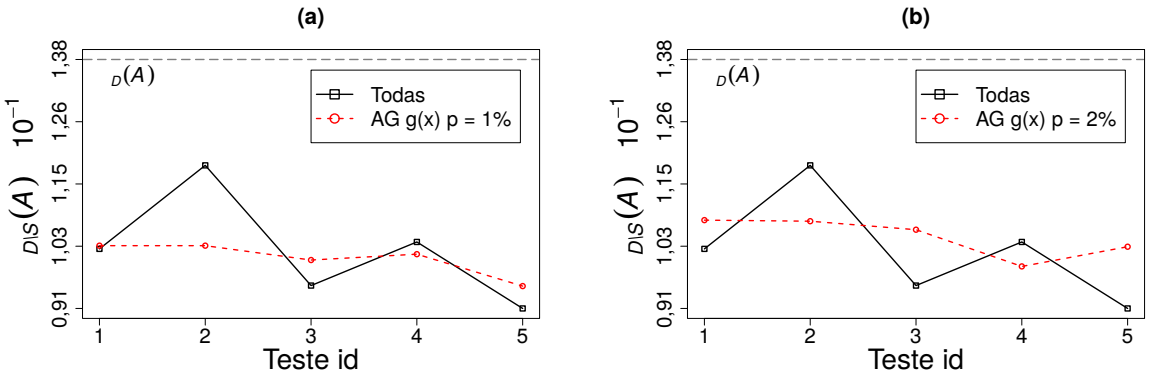

Figura 4.13: Instabilidade resultante após a remoção de amostras instáveis que preservam a estrutura da base de dados *Wine* utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a segunda função de avaliação (Equação 3.4) com $p = 1\%$ (a) e $p = 2\%$ (b). A remoção de todas a amostras apresentou melhores resultados de instabilidade em 60% dos testes (testes 1, 3 e 5, para ambos os valores de p) e a seleção e remoção utilizando algoritmo genético apresentou melhores resultados em 40% dos testes (testes 2 e 4, para ambos os valores de p). A linha tracejada representa a instabilidade ($\mu_D(A)$) da base de dados sem remoção de amostras.

pela remoção de todas as amostras e pela seleção e remoção utilizando algoritmo genético, com $p = 1\%$ está no intervalo $-0{,}04 \times 10^{-1}$ (testes 3 e 5) a $0{,}15 \times 10^{-1}$ (teste 2) e com p

Tabela 4.18: Instabilidade resultante após a remoção de amostras instáveis que preservam a estrutura da base de dados *Wine* utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a segunda função de avaliação (Equação 3.4). A remoção de todas a amostras apresentou melhores resultados de instabilidade em 60% dos testes (testes 1, 3 e 5, para ambos os valores de p) e a seleção e remoção utilizando algoritmo genético apresentou melhores resultados em 40% dos testes (testes 2 e 4, para ambos os valores de p).

p	k	c	Teste id	Todas	AG $g(x)$	Todas − AG $g(x)$
1%	2	1,750	1	$\mathbf{1{,}02 \times 10^{-1}}$	$1{,}03 \times 10^{-1}$	$-0{,}01 \times 10^{-1}$
		1,922	2	$1{,}18 \times 10^{-1}$	$\mathbf{1{,}03 \times 10^{-1}}$	$0{,}15 \times 10^{-1}$
		1,849	3	$\mathbf{0{,}96 \times 10^{-1}}$	$1{,}00 \times 10^{-1}$	$-0{,}04 \times 10^{-1}$
		1,849	4	$1{,}04 \times 10^{-1}$	$\mathbf{1{,}01 \times 10^{-1}}$	$0{,}03 \times 10^{-1}$
		1,485	5	$\mathbf{0{,}91 \times 10^{-1}}$	$0{,}95 \times 10^{-1}$	$-0{,}04 \times 10^{-1}$
2%	2	0,857	1	$\mathbf{1{,}02 \times 10^{-1}}$	$1{,}08 \times 10^{-1}$	$-0{,}06 \times 10^{-1}$
		0,941	2	$1{,}18 \times 10^{-1}$	$\mathbf{1{,}08 \times 10^{-1}}$	$0{,}10 \times 10^{-1}$
		0,906	3	$\mathbf{0{,}96 \times 10^{-1}}$	$1{,}06 \times 10^{-1}$	$-0{,}10 \times 10^{-1}$
		0,906	4	$1{,}04 \times 10^{-1}$	$\mathbf{0{,}99 \times 10^{-1}}$	$0{,}05 \times 10^{-1}$
		0,727	5	$\mathbf{0{,}91 \times 10^{-1}}$	$1{,}03 \times 10^{-1}$	$-0{,}12 \times 10^{-1}$

= 2% está no intervalo $-0{,}12 \times 10^{-1}$ (teste 5) a $0{,}10 \times 10^{-1}$ (teste 2).

Uma vez que se conhece a instabilidade resultante após a remoção de amostras instáveis que preservam a estrutura dos dados utilizando ambas as abordagens, a próxima análise avalia a distribuição das amostras removidas entre os grupos com o propósito de verificar se foi possível reduzir o número de amostras removidas e quanto essa redução afetou a instabilidade resultante.

4.3.3 DISTRIBUIÇÃO DAS AMOSTRAS REMOVIDAS

As Tabelas 4.19 , 4.20 e 4.21 apresentam a distribuição das amostras instáveis que preservam a estrutura dos dados removidas entre os grupos da base de dados *Wine* utilizando os seguintes critérios:

i) identificação e remoção de todas as amostras instáveis (Equação 2.14, pg. 31) que isoladamente preservam a estrutura dos dados (Equação 2.17, pg. 32), proposto por Mulder (2014);

ii) seleção e remoção proposta neste trabalho, no qual um subconjunto de amostras instáveis (Equação 2.14, pg. 31) que preserva a estrutura dos dados (Equação 3.2, pg. 48) é selecionado e removido utilizando algoritmo genético com duas funções de avaliação.

A Tabela 4.19 apresenta os resultados obtidos utilizando a primeira função de avaliação proposta para o algoritmo genético (Equação 3.3, pg. 49).

Tabela 4.19: Distribuição das amostras instáveis que preservam a estrutura dos dados selecionadas e removidas entre os grupos da base de dados *Wine* utilizando a primeira função de avaliação (Equação 3.3). A seleção e remoção de amostras utilizando algoritmo genético apresentou redução no percentual de amostras removidas em todos os testes.

Grupo	Teste id	Todas	AG $f(x)$	Todas − AG $f(x)$
tipo 1	1	7 (11,86%)	**6 (10,17%)**	1 (1,69%)
tipo 2		19 (26,76%)	**9 (12,68%)**	10 (14,08%)
tipo 3		30 (62,5%)	**16 (33,33%)**	14 (29,17%)
tipo 1	2	6 (10,17%)	**4 (6,78%)**	2 (3,39%)
tipo 2		18 (25,35%)	**9 (12,68%)**	9 (12,67%)
tipo 3		27 (56,25%)	**14 (29,17%)**	13 (27,08%)
tipo 1	3	5 (8,47%)	**2 (3,39%)**	3 (5,08%)
tipo 2		18 (25,35%)	**10 (14,08%)**	8 (11,27%)
tipo 3		30 (62,5%)	**14 (29,17%)**	16 (33,33%)
tipo 1	4	8 (13,56%)	**3 (5,08%)**	5 (8,48%)
tipo 2		19 (26,76%)	**12 (16,9%)**	7 (9,86%)
tipo 3		26 (54,17%)	**13 (27,08%)**	13 (27,09%)
tipo 1	5	9 (15,25%)	**4 (6,78%)**	5 (8,47%)
tipo 2		24 (33,8%)	**11 (15,49%)**	13 (18,31%)
tipo 3		33 (68,75%)	**17 (35,42%)**	16 (33,33%)

Considerando a distribuição das amostras instáveis que preservam a estrutura dos dados selecionadas e removidas utilizando a primeira função de avaliação (Tabela 4.19), é possível notar que utilizando algoritmo genético em todos os testes foi possível reduzir o número de amostras removidas em relação à estratégia de remoção de todas as amostras. Quanto ao percentual de amostras removidas no grupo *tipo 1*, a redução alcançada está entre 1,69% (teste 1) e 8,48% (teste 4), para o grupo *tipo 2* a redução está entre 9,86% (teste 4) e 18,31% (teste 5) e no grupo *tipo 3* a redução alcançada está entre 27,08% (teste

2) e 33,33% (testes 3 e 5). No melhor caso, para o grupo *tipo 1* (teste 4), foi possível reduzir o percentual de 13,56% para 5,08%, para o grupo *tipo 2* (teste 5), foi possível reduzir o percentual de 33,8% para 15,49% e para o grupo *tipo 3*, foi possível reduzir o percentual de 62,5% para 29,17% (teste 3) e 68,75% para 35,42% (teste 5).

Analisando a instabilidade resultante após a seleção e remoção de amostras utilizando a primeira função de avaliação (Tabela 4.17 e Figura 4.12) em relação à distribuição dessas amostras (Tabela 4.19), nota-se que com reduções entre $0{,}07 \times 10^{-1}$ (testes 1 e 4) e $0{,}21 \times 10^{-1}$ (teste 2) – apesar de aumentos entre $0{,}06 \times 10^{-1}$ (teste 5) e $0{,}07 \times 10^{-1}$ (teste 3) – na instabilidade apresentada pela remoção de todas as amostras, a abordagem proposta com algoritmo genético conseguiu reduções de 1,69% a 8,48% no percentual de amostras removidas no grupo *tipo 1*, reduções de 9,86% a 18,31% no percentual de amostras removidas no grupo *tipo 2* e reduções de 27,08% a 33,33% no percentual de amostras removidas no grupo *tipo 3*.

A Tabela 4.20 apresenta os resultados obtidos utilizando a segunda função de avaliação (Equação 3.4, pg. 49) com o percentual de redução p igual a 1%, e o coeficiente k igual a 2.

Considerando a distribuição das amostras instáveis que preservam a estrutura dos dados selecionadas e removidas utilizando a segunda função de avaliação e $p = 1\%$ (Tabela 4.20), é possível notar que utilizando algoritmo genético em todos os testes foi possível reduzir o número de amostras removidas em relação à estratégia de remoção de todas as amostras. Quanto ao percentual de amostras removidas no grupo *tipo 1*, a redução alcançada está entre 1,69% (teste 3) e 8,48% (teste 4), para o grupo *tipo 2* a redução alcançada está entre 8,45% (teste 4) e 21,12% (teste 5) e para o grupo *tipo 3* a redução alcançada está entre 27,09% (teste 4) e 43,75% (teste 3). No melhor caso, para o grupo *tipo 1* (teste 4), foi possível reduzir o percentual de 13,56% para 5,08%, para o grupo *tipo 2* (teste 5), foi possível reduzir o percentual de 33,8% para 12,68% e para o grupo *tipo 3* (teste 3) foi possível reduzir o percentual de 62,5% para 18,75%.

Analisando a instabilidade resultante após a seleção e remoção de amostras utilizando a segunda função de avaliação e $p = 1\%$ (Tabela 4.18 e Figura 4.13) em relação à distribuição dessas amostras (Tabela 4.20), nota-se que com reduções entre $0{,}03 \times 10^{-1}$ (teste 4) e $0{,}15 \times 10^{-1}$ (teste 2) – apesar de aumentos entre $0{,}01 \times 10^{-1}$ (teste 1) e $0{,}04 \times 10^{-1}$ (testes 3 e 5) – na instabilidade apresentada pela remoção de todas as amostras, a abordagem

Tabela 4.20: Distribuição das amostras instáveis que preservam a estrutura dos dados selecionadas e removidas entre os grupos da base de dados *Wine* utilizando a segunda função de avaliação com $p = 1\%$ e $k = 2$ (Equação 3.4). A seleção e remoção de amostras utilizando algoritmo genético apresentou redução no percentual de amostras removidas em todos os testes.

Grupo	Teste id	Todas	AG $g(x)$ $p = 1\%$	Todas − AG $g(x)$
tipo 1	1	7 (11,86%)	**3 (5,08%)**	4 (6,78%)
tipo 2		19 (26,76%)	**8 (11,27%)**	11 (15,49%)
tipo 3		30 (62,5%)	**13 (27,08%)**	17 (35,42%)
tipo 1	2	6 (10,17%)	**3 (5,08%)**	3 (5,09%)
tipo 2		18 (25,35%)	**8 (11,27%)**	10 (14,08%)
tipo 3		27 (56,25%)	**12 (25%)**	15 (31,25%)
tipo 1	3	5 (8,47%)	**4 (6,78%)**	1 (1,69%)
tipo 2		18 (25,35%)	**9 (12,68%)**	9 (12,67%)
tipo 3		30 (62,5%)	**9 (18,75%)**	21 (43,75%)
tipo 1	4	8 (13,56%)	**3 (5,08%)**	5 (8,48%)
tipo 2		19 (26,76%)	**13 (18,31%)**	6 (8,45%)
tipo 3		26 (54,17%)	**13 (27,08%)**	13 (27,09%)
tipo 1	5	9 (15,25%)	**4 (6,78%)**	5 (8,47%)
tipo 2		24 (33,8%)	**9 (12,68%)**	15 (21,12%)
tipo 3		33 (68,75%)	**18 (37,5%)**	15 (31,25%)

proposta com algoritmo genético conseguiu reduções de 1,69% a 8,48% no percentual de amostras removidas no grupo *tipo 1*, reduções de 8,45% a 21,12% no percentual de amostras removidas no grupo *tipo 2* e reduções de 27,09% a 43,75% no percentual de amostras removidas no grupo *tipo 3*. Isto é, utilizando a segunda função de avaliação e definindo que o número de amostras para remoção n só poderia ser multiplicado por 2 ($k = 2$) caso houvesse redução de pelo menos 1% ($p = 1\%$) da instabilidade resultante, foi possível alcançar reduções de 1,69% a 8,48% no percentual de amostras removidas no grupo *tipo 1*, reduções de 8,45% a 21,12% no percentual de amostras removidas no grupo *tipo 2* e reduções de 27,09% a 43,75% no percentual de amostras removidas no grupo *tipo 3*.

A Tabela 4.21 apresenta os resultados obtidos utilizando ainda a segunda função de avaliação, com o percentual de redução p igual a 2% e o coeficiente k igual a 2.

Tabela 4.21: Distribuição das amostras instáveis que preservam a estrutura dos dados selecionadas e removidas entre os grupos da base de dados *Wine* utilizando a segunda função de avaliação com $p = 2\%$ e $k = 2$ (Equação 3.4). A seleção e remoção de amostras utilizando algoritmo genético apresentou considerável redução no percentual de amostras removidas em todos os testes.

Grupo	Teste id	Todas	AG $g(x)$ $p = 2\%$	Todas − AG $g(x)$
tipo 1	1	7 (11,86%)	**0**	7 (11,86%)
tipo 2		19 (26,76%)	**5 (7,04%)**	14 (19,72%)
tipo 3		30 (62,5%)	**5 (10,42%)**	25 (52,08%)
tipo 1	2	6 (10,17%)	**2 (3,39%)**	4 (6,78%)
tipo 2		18 (25,35%)	**2 (2,82%)**	16 (22,53%)
tipo 3		27 (56,25%)	**5 (10,42%)**	22 (45,83%)
tipo 1	3	5 (8,47%)	**1 (1,69%)**	4 (6,78%)
tipo 2		18 (25,35%)	**5 (7,04%)**	13 (18,31%)
tipo 3		30 (62,5%)	**7 (14,58%)**	23 (47,92%)
tipo 1	4	8 (13,56%)	**4 (6,78%)**	4 (6,78%)
tipo 2		19 (26,76%)	**7 (9,86%)**	12 (16,90%)
tipo 3		26 (54,17%)	**6 (12,5%)**	20 (41,67%)
tipo 1	5	9 (15,25%)	**1 (1,69%)**	8 (13,56%)
tipo 2		24 (33,8%)	**8 (11,27%)**	16 (22,53%)
tipo 3		33 (68,75%)	**9 (18,75%)**	24 (50,00%)

Considerando a distribuição das amostras instáveis que preservam a estrutura dos dados selecionadas e removidas utilizando a segunda função de avaliação e *p = 2%* (Tabela 4.21), é possível notar que utilizando algoritmo genético em todos os testes foi possível reduzir consideravelmente o número de amostras removidas em relação à estratégia de remoção de todas as amostras. Quanto ao percentual de amostras removidas no grupo *tipo 1*, a redução alcançada está entre 6,78% (testes 2, 3 e 4) e 13,56% (teste 5), para o grupo *tipo 2* a redução alcançada está entre 16,9% (teste 4) e 22,53% (testes 2 e 5) e para o grupo *tipo 3* a redução alcançada está entre 41,67% (teste 4) e 52,08% (teste 1). No melhor caso, para o grupo *tipo 1* (teste 5), foi possível reduzir o percentual de 15,25% para 1,69%, para o grupo *tipo 2*, foi possível reduzir o percentual de 25,35% para 2,82% (teste 2) e 33,8% para 11,27% (teste 5) e para o grupo *tipo 3* foi possível reduzir o percentual de 62,5% para 10,42%.

Analisando a instabilidade resultante após a seleção e remoção de amostras utilizando a segunda função de avaliação e $p = 2\%$ (Tabela 4.18 e Figura 4.13) em relação à distribuição dessas amostras (Tabela 4.21), nota-se que com reduções entre $0{,}05 \times 10^{-1}$ (teste 4) e $0{,}10 \times 10^{-1}$ (teste 2) – apesar de aumentos entre $0{,}06 \times 10^{-1}$ (teste 1) e $0{,}12 \times 10^{-1}$ (teste 5) – na instabilidade apresentada pela remoção de todas amostras, a abordagem proposta com algoritmo genético conseguiu reduções de 6,78% a 13,56% no percentual de amostras removidas no grupo *tipo 1*, reduções de 16,9% a 22,53% no percentual de amostras removidas no grupo *tipo 2* e reduções de 41,67% a 52,08% no percentual de amostras removidas no grupo *tipo 3*. Isto é, definindo que para a segunda função de avaliação, o número de amostras para remoção n só poderia ser multiplicado por 2 ($k = 2$) caso houvesse redução de pelo menos 2% ($p = 2\%$) da instabilidade apresentada em casos nos quais apenas n amostras fossem removidas, foi possível alcançar reduções de 6,78% a 13,56% no percentual de amostras removidas no grupo *tipo 1*, reduções de 16,9% a 22,53% no percentual de amostras removidas no grupo *tipo 2* e reduções de 41,67% a 52,08% no percentual de amostras removidas no grupo *tipo 3*.

5 CONSIDERAÇÕES FINAIS

Neste capítulo são apresentadas as considerações finais deste trabalho. Inicialmente, é apresentada uma discussão geral sobre os resultados de agrupamento e análises de estabilidade considerando todos os ambientes de testes – apresentados no Capítulo 4. Logo após, são apresentadas as conclusões, contribuições, limitações e possibilidades de trabalhos futuros.

5.1 DISCUSSÃO GERAL

Nesta seção são apresentadas uma discussão geral sobre os resultados de agrupamento obtidos pela aplicação dos algoritmos HDDC, SOM e DBSCAN sobre a base de dados de SNPs e uma avaliação geral das análises de estabilidade do agrupamento e do conjunto de dados, realizadas utilizando a abordagem de Mulder (2014) e a abordagem proposta neste trabalho.

5.1.1 AGRUPAMENTO DA BASE DE DADOS DE SNPS

Por meio da aplicação dos algoritmos HDDC, SOM e DBSCAN sobre a base de dados de SNPs, foi possível adquirir e validar alguns conhecimentos quanto ao relacionamento das amostras e estrutura da base de dados.

Considerando a avaliação dos resultados de agrupamento, houve concordância entre os algoritmos quanto à formação de dois, três ou quatro grupos. Nos casos em que dois grupos foram formados, um grupo foi composto pela junção de amostras das raças Holandesa e Jersey e outro grupo foi composto pelas amostras da raça Nelore (Figura 4.2, instância 1, pg. 57). Nos casos em que três grupos foram formados, houve duas situações:

i) um grupo foi composto pelas amostras da raça Nelore, um segundo grupo foi dedicado para as amostras da raça Holandesa e um baixo percentual de amostras da raça Jersey e um terceiro grupo foi composto por alto percentual de amostras da raça Jersey (Figura 4.2, instância 2, pg. 57);

ii) um grupo foi composto pelas amostras da raça Nelore, um segundo grupo foi composto por um baixo percentual de amostras da raça Jersey e um terceiro grupo foi composto

por altos percentuais de amostras das raças Holandesa e Jersey (Figura 4.4, instâncias 2, 4 e 6, pg. 60).

Nos casos em quatro grupos foram formados, também houve duas situações:

i) um grupo foi composto pelas amostras da raça Nelore, um segundo grupo foi composto pelas amostras da raça Holandesa, um terceiro grupo foi composto por um alto percentual de amostras da raça Jersey e um quarto grupo foi composto por um baixo percentual de amostras da raça Jersey (Figura 4.4, instâncias 1, 3 e 5, pg. 60);

ii) um grupo foi composto pelas amostras da raça Nelore, um segundo grupo foi composto por um alto percentual de amostras da raça Jersey, um terceiro grupo foi composto por um alto percentual de amostras da raça Holandesa e um quarto grupo foi composto por baixos percentuais de amostras das raças Holandesa e Jersey (Figura 4.6, exceto instância 1, pg. 62).

De modo geral, tais resultados de agrupamento apresentam características como a formação de grupos exclusivos para as amostras da raça Nelore, a formação de grupos dedicados para cada raça taurina (Holandesa e Jersey), a formação de grupos compostos pela junção de amostras das raças Holandesa e Jersey e a divisão das amostras de determinada raça em subgrupos. Estas características validam o conhecimento quanto ao número de grupos da base de dados e validam também as hipóteses quanto à semelhança entre as raças taurinas (Holandesa e Jersey) e a dissimilaridade entre tais raças e a raça zebuína (Nelore). Ver Figura 4.1 (pg. 56) para mais detalhes sobre o resultado de agrupamento esperado. A subdivisão dentro de determinada raça pode ser um indicativo de que mesmo pertencendo à mesma raça, algumas amostras podem ser separadas de acordo com um conjunto de marcadores ou características.

Considerando os resultados de agrupamento apresentados pelo algoritmo DBSCAN (Figura 4.6, pg. 62), algoritmo capaz de identificar amostras como ruídos, pode-se atribuir confiabilidade à base de dados. No pior caso, apenas 2% das amostras da raça Holandesa são identificadas como ruídos. Tal percentual de ruídos atribui confiabilidade à base de dados sugerindo, por exemplo, que a mesma não foi prejudicada por possíveis erros durante a coleta das amostras.

Considerando as avaliações dos resultados de agrupamento com os índices CH e DI, é possível notar que houve concordância entre tais índices para os resultados de agrupa-

mento apresentados pelos algoritmos HDDC e DBSCAN (Tabelas 4.1 e 4.3, pg. 59 e 64, respectivamente), de modo que não houve concordância apenas para os resultados de agrupamento apresentados pelo algoritmo SOM (Tabela 4.2, pg. 61).

5.1.2 SELEÇÃO E REMOÇÃO DE AMOSTRAS PREJUDICIAIS

Considerando a seleção e remoção de amostras prejudiciais, observa-se que em todos os testes, para todas as bases de dados, foi possível reduzir a instabilidade original da base de dados por meio da aplicação do algoritmo genético, independentemente da função de avaliação utilizada. Isto é, em nenhum teste a abordagem aqui proposta utilizando algoritmo genético apresentou aumento de instabilidade após a seleção e remoção de amostras prejudiciais aos resultados de agrupamento que deveriam ser estudadas isoladamente.

Por outro lado, a remoção de todas as amostras prejudiciais, sugerida por Mulder (2014), apresentou considerável aumento de instabilidade (de $7{,}28 \times 10^{-4}$ para $19{,}81 \times 10^{-4}$, ver Tabelas 4.5 e 4.6, pg. 66 e 68, respectivamente) para a base de dados de SNPs. Esse fato confirma que a remoção de amostras que foram avaliadas individualmente – quanto à preservação de estrutura – pode prejudicar a estrutura dos dados, implicando em aumento de instabilidade, como descrito nas Seções 1.1 e 3.5.1.

Analisando a instabilidade resultante após a seleção e remoção de amostras instáveis que preservam a estrutura das bases de dados utilizando algoritmo genético, ilustrada pelas Figuras 5.1, 5.2 e 5.3, observa-se que para a maioria dos casos, utilizando a primeira função de avaliação (Equação 3.3, pg. 49, aqui representada por AG f(x)), o algoritmo genético alcança melhores resultados do que quando se utiliza a segunda função de avaliação (Equação 3.4, pg. 49, aqui representada por AG g(x)). O que justifica esse comportamento é o fato de que a primeira função de avaliação, $f(x)$, ao contrário da segunda, não sofre nenhuma restrição quanto ao número de amostras removidas. Dessa forma, o algoritmo genético utilizando $f(x)$ pode remover quantas amostras prejudiciais forem necessárias para minimizar a instabilidade resultante.

Quanto à influência do percentual (p) de redução esperado na instabilidade resultante sobre a segunda função de avaliação, $g(x)$, constata-se que para a maioria dos casos ilustrados pelas Figuras 5.1, 5.2 e 5.3, utilizando $p = 1\%$, o algoritmo genético alcança melhores resultados de instabilidade que quando utilizando $p = 2\%$. Esse comportamento se justifica pela análise do fator c, apresentada na Seção 3.5.2 e ilustrada pela Figura 3.1

(pg. 51), na qual verifica-se que para o mesmo coeficiente k, quanto menor o valor p, mais relaxada será a relação entre a instabilidade resultante e o número de amostras removidas, o que atribui maior prioridade para a redução de instabilidade. Desse modo, quando p é definido como 1%, o algoritmo genético recebe menor restrição quanto ao número de amostras e, portanto, melhores resultados de instabilidade podem ser alcançados.

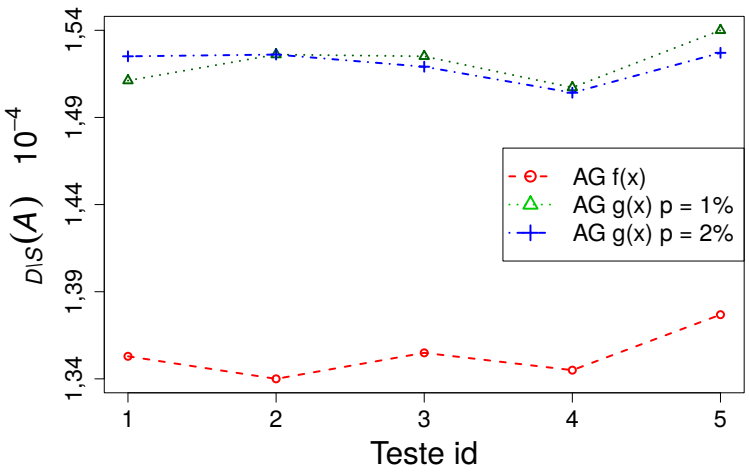

Figura 5.1: Instabilidade resultante após a seleção e remoção de amostras instáveis que preservam a estrutura da base de dados de SNPs utilizando o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a primeira e a segunda função de avaliação (Equações 3.3 e 3.4, $f(x)$ e $g(x)$, respectivamente). A instabilidade original (sem remoção de amostras) é $7{,}28 \times 10^{-4}$ e a instabilidade resultante após a remoção de todas as amostras é $19{,}81 \times 10^{-4}$.

5.1.3 DISTRIBUIÇÃO DAS AMOSTRAS REMOVIDAS

A Figura 5.4 ilustra a distribuição das amostras instáveis que preservam a estrutura da base de dados de SNPs dentro da raça Holandesa removidas por cada abordagem e a Figura 5.5 ilustra a distribuição das amostras instáveis que preservam a estrutura da base de dados *Wine* dentro do grupo *tipo 3* removidas por cada abordagem. Ambas apresentam tal distribuição em relação à instabilidade resultante após a remoção dessas amostras.

Considerando a distribuição das amostras removidas, ilustrada pelas Figuras 5.4 e 5.5, constata-se que em todos os testes, para todas as bases de dados, foi possível reduzir o número de amostras removidas por meio da aplicação do algoritmo genético, independentemente da função de avaliação utilizada, evitando assim, remoções excessivas de amostras.

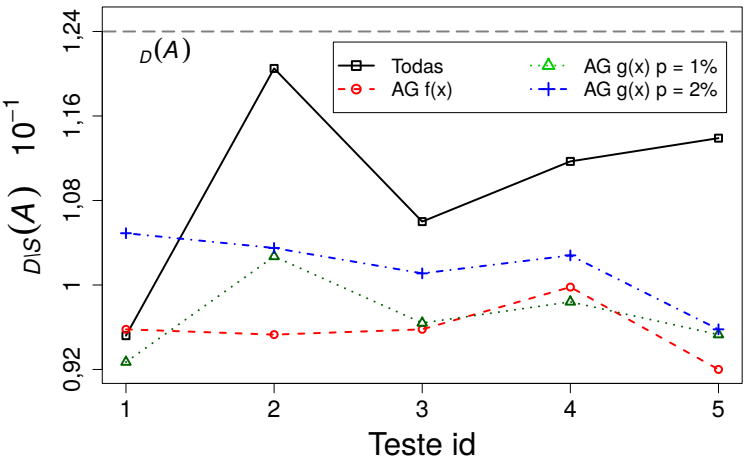

Figura 5.2: Instabilidade resultante após a remoção de amostras instáveis que preservam a estrutura da base de dados *Iris* utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a primeira e a segunda função de avaliação (Equações 3.3 e 3.4, $f(x)$ e $g(x)$, respectivamente). A instabilidade original (sem remoção de amostras), representada pela linha tracejada ($\mu_D(A)$), é $1{,}24 \times 10^{-1}$.

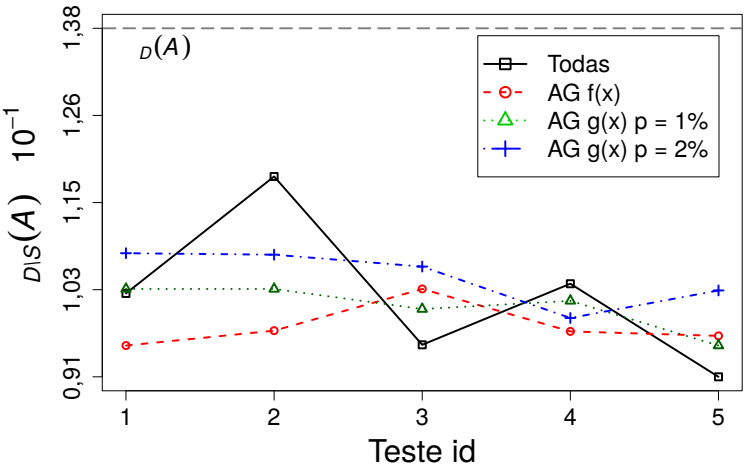

Figura 5.3: Instabilidade resultante após a remoção de amostras instáveis que preservam a estrutura da base de dados *Wine* utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a primeira e a segunda função de avaliação (Equações 3.3 e 3.4, $f(x)$ e $g(x)$, respectivamente). A instabilidade original (sem remoção de amostras), representada pela linha tracejada ($\mu_D(A)$), é $1{,}38 \times 10^{-1}$.

Utilizando a abordagem de remoção de todas as amostras, sugerida por Mulder (2014), na base de dados de SNPs 84,06% das amostras da raça Holandesa seriam removidas

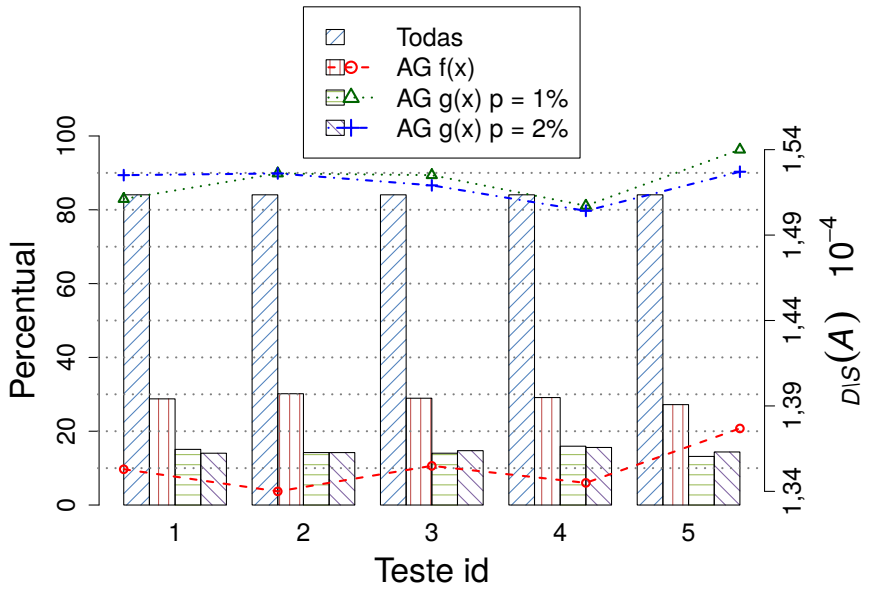

Figura 5.4: Distribuição das amostras instáveis que preservam a estrutura dos dados dentro da raça Holandesa da base de dados de SNPs removidas utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a primeira e a segunda função de avaliação (Equações 3.3 e 3.4, $f(x)$ e $g(x)$, respectivamente). A instabilidade original (sem remoção de amostras) é $7{,}28 \times 10^{-4}$ e a instabilidade resultante após a remoção de todas as amostras é $19{,}81 \times 10^{-4}$.

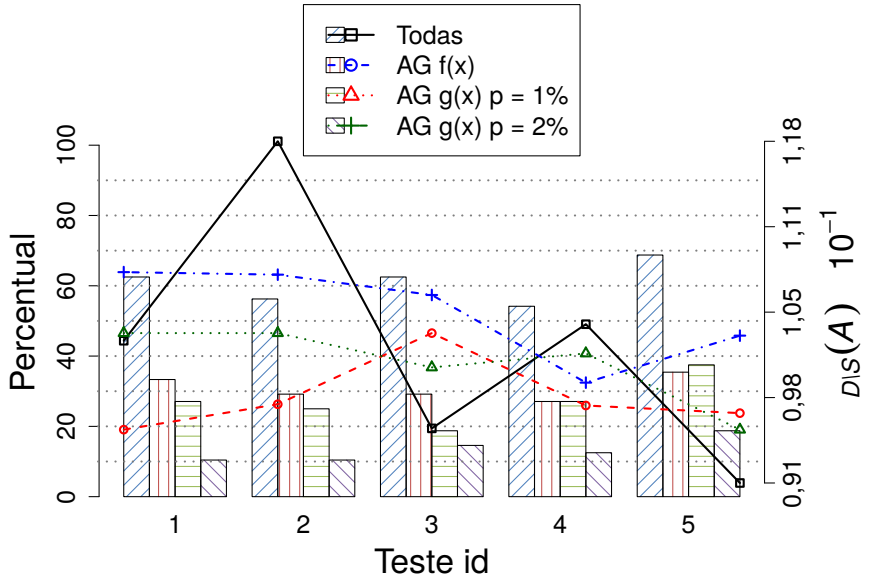

Figura 5.5: Distribuição das amostras instáveis que preservam a estrutura dos dados dentro do grupo *tipo 3* da base de dados *Wine* removidas utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a primeira e a segunda função de avaliação (Equações 3.3 e 3.4, $f(x)$ e $g(x)$, respectivamente). A instabilidade original (sem remoção de amostras) é $1{,}38 \times 10^{-1}$.

(Figura 5.4) e na base de dados *Wine*, no pior caso, 68,75% das amostras do grupo *tipo 3* seriam removidas (Figura 5.5, teste 5).

Todavia, utilizando algoritmo genético e a primeira função de avaliação (Equação 3.3, pg. 49, aqui representada por AG f(x)) que não sofre nenhuma restrição quanto ao número de amostras removidas, na base de dados de SNPs, no pior caso, 30,16% das amostras da raça Holandesa seriam removidas (Figura 5.4, teste 2) e na base de dados *Wine*, no pior caso, 35,42% das amostras do grupo *tipo 3* seriam removidas (Figura 5.5, teste 5). Isto é, mesmo utilizando a função de avaliação que não sofre nenhuma restrição quanto ao número de amostras removidas, o algoritmo genético conseguiu consideráveis reduções no percentual de amostras removidas, evitando assim remoções excessivas de amostras.

Analisando a redução alcançada no percentual de amostras selecionadas e removidas em relação à função de avaliação utilizada pelo algoritmo genético, constata-se que a primeira função de avaliação apresenta o maior percentual de amostras removidas em comparação com o percentual apresentado pela segunda função, uma vez que, como citado anteriormente, a primeira função de avaliação não sofre nenhuma restrição quanto ao número de amostras removidas. Considerando a segunda função de avaliação, observa-se que existe uma tendência de que quanto maior o valor de p, menor o percentual de amostras removidas, assim como é discutido pela análise do fator c, apresentada na Seção 3.5.2 e ilustrada pela Figura 3.1 (pg. 51).

Considerando a distribuição das amostras removidas em relação à instabilidade resultante após a remoção de tais amostras (Figuras 5.4 e 5.5), nota-se que em todos os testes, para todas as bases de dados, mesmo reduzindo o percentual de amostras removidas, o algoritmo genético conseguiu reduzir a instabilidade original da base de dados, independentemente da função de avaliação utilizada. Isto é, em nenhum teste a abordagem aqui proposta utilizando algoritmo genético apresentou reduções no percentual de amostras removidas que aumentassem a instabilidade original da base de dados, de modo que todas as reduções nos percentuais de amostras removidas garantiram redução da instabilidade original.

A distribuição das amostras instáveis que preservam a estrutura da base de dados de SNPs dentro da raça Holandesa removidas por cada abordagem é ilustrada pela Figura 5.4 e a distribuição das amostras instáveis que preservam a estrutura da base de dados *Wine* dentro do grupo *tipo 3* é ilustrada pela Figura 5.5. As distribuições de tais amostras para

os demais grupos da base de dados de SNPs são ilustradas no Apêndice A, para os demais grupos da base de dados *Wine* são ilustradas no Apêndice B e para a base de dados *Iris* são ilustradas no Apêndice C.

5.2 CONCLUSÃO

Considerando a discussão apresentada anteriormente, nota-se que por meio de ajustes no conceito de amostras prejudiciais aos resultados de agrupamento e do acréscimo de algoritmo genético para seleção de tais amostras, a abordagem aqui proposta é capaz de solucionar características indesejáveis – identificadas na abordagem de Mulder et al. (2010) e Mulder (2014) – que podem resultar em remoção excessiva de amostras e ainda não garantem melhoria de estabilidade. Além de solucionar tais questões, a abordagem aqui proposta também permite que o usuário controle o processo de análise, o que atribui maior aplicabilidade e confiabilidade para tal processo. É importante destacar que mesmo com tais ajustes a abordagem aqui proposta preserva a independência do processo de análise quanto ao algoritmo de agrupamento utilizado – característica importante apresentada pela abordagem de Mulder et al. (2010) e Mulder (2014) que permite a utilização de outros algoritmos de agrupamento.

Tais contribuições são alcançadas principalmente pela utilização de algoritmo genético – que trata a identificação de amostras prejudiciais como um problema de otimização – com funções de avaliação que buscam minimizar a instabilidade resultante e, ao mesmo tempo, o número de amostras removidas – como a segunda função de avaliação, $g(x)$, representada pela Equação 3.4 (pg. 49). Ainda considerando tal discussão, observa-se que a abordagem aqui proposta pode ser considerada como uma ferramenta promissora para aquisição e validação de conhecimento em bases de dados genotípicos.

Assim como a abordagem de Mulder et al. (2010) e Mulder (2014), a abordagem aqui proposta é não-supervisionada, isto é, utiliza amostras não-rotuladas. Dessa forma, durante o processo de seleção das amostras para remoção o algoritmo genético – utilizando a segunda função de avaliação – avalia o número total de amostras removidas em relação ao número total de amostras da base de dados. Para aplicações nas quais as amostras são rotuladas, tal característica pode ser considerada como uma limitação, uma vez que conhecendo os grupos, poderia ser definida uma função de avaliação para o algoritmo genético que considerasse o número de amostras removidas por cada grupo.

Como possibilidades de trabalhos futuros, quanto à utilização da base de dados genotípicos, existem diversas análises que podem ser realizadas a partir dos resultados de agrupamento e estabilidade aqui apresentados. Mais especificamente, sugere-se analisar as subdivisões de amostras da mesma raça. Tais subdivisões podem ser consideradas como indicativos de que mesmo pertencendo à mesma raça, algumas amostras podem ser separadas por um conjunto de marcadores. Dessa forma, sugere-se analisar tais subdivisões com o intuito de encontrar tal conjunto de marcadores que distingue essas amostras.

Ainda quanto à utilização da base de dados genotípicos, sugere-se também analisar as amostras instáveis que preservam a estrutura dos dados. Como citado anteriormente, tais amostras podem ser consideradas como notáveis e deveriam ser estudadas isoladamente uma vez que fogem ao comportamento geral do conjunto de dados, ou seja, não podem ser agrupadas adequadamente. Por último, sugere-se analisar as subdivisões de amostras da mesma raça e as amostras notáveis com relação à outros dados fenotípicos, como a produção de leite, por exemplo. Além de auxiliar as análises de subdivisões e amostras notáveis, a utilização de outros dados fenotípicos pode resultar em maior aquisição de conhecimento com potencial econômico e científico.

Em relação ao processo de análise, como possibilidades de trabalhos futuros, sugere-se analisar a abordagem aqui proposta em outras bases de dados, a utilização de outros algoritmos de agrupamento e a realização de testes com variações nos parâmetros p e k para o fator c (Equação 3.5, pg. 50) da segunda função de avaliação do algoritmo genético (Equação 3.4, pg. 49). O desenvolvimento de outras funções de avaliação para o algoritmo genético pode ser considerado como outra possibilidade de trabalho futuro. Mais especificamente, pode-se desenvolver funções de avaliação para aplicações que utilizam amostras rotuladas, como citado anteriormente. Outra possibilidade seria analisar a utilização de outras técnicas de otimização além de algoritmos genéticos.

Por último, sugere-se analisar a utilização de técnicas de redução de dimensionalidade em junção com a abordagem aqui proposta. A utilização de tais técnicas além de reduzir o custo computacional podem auxiliar a análise de estabilidade dos dados e as análises – citadas anteriormente – com dados fenotípicos, destacando, por exemplo, um conjunto de marcadores fortemente relacionado com determinada característica física, fisiológica ou morfológica.

REFERÊNCIAS

ARTERO, A. O. **Inteligência Artificial: teórica e prática**. São Paulo: Editora Livraria da Fisica, 2009.

BERGÉ, L.; BOUVEYRON, C.; GIRARD, S. HDclassif: An R package for model-based clustering and discriminant analysis of high-dimensional data. **Journal of Statistical Software**, v. 46, n. 6, p. 1–29, 2012.

BISHOP, C. M. **Pattern recognition and machine learning**. New York: Springer New York, 2006.

BOLDRINI, J. L.; COSTA, S. I. R.; FIGUEREDO, V. L.; WETZLER, H. G. **Álgebra linear**. 3. ed. São Paulo: Harper & Row do Brasil, 1980.

BOUVEYRON, C.; GIRARD, S.; SCHMID, C. High-dimensional data clustering. **Computational Statistics & Data Analysis**, v. 52, n. 1, p. 502–519, 2007.

CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics-theory and Methods**, v. 3, n. 1, p. 1–27, 1974.

DARWIN, C. **On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life**. London: John Murray, 1859.

DEMPSTER, A.; LAIRD, N.; RUBIN, D. Maximum likelihood from incomplete data via the em algorithm. **Journal of the Royal Statistical Society**, v. 39, n. 1, p. 1–38, 1977.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. 2nd. ed. Danvers: John Wiley & Sons, 2001.

DUNN, J. C. Well-separated clusters and optimal fuzzy partitions. **Journal of cybernetics**, v. 4, n. 1, p. 95–104, 1974.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)**, 1996. p. 226–231.

FIELDING, A. H. **Cluster and classification techniques for the biosciences**. New York: Cambridge University Press, 2007.

FORGY, E. W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. **Biometrics**, v. 21, p. 768–769, 1965.

GHR. **What are genome-wide association studies?**, 2015. Disponível em: <http://ghr.nlm.nih.gov/handbook/genomicresearch/gwastudies>.

GHR. **What are single nucleotide polymorphism (SNPs)?**, 2015. Disponível em: <http://ghr.nlm.nih.gov/handbook/genomicresearch/snp>.

GOLDBERG, D. E. **Genetic algorithms in search, optimization, and machine learning**. Boston: Addison-Wesley, 1989.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. **Journal of Intelligent Information Systems**, v. 17, n. 2-3, p. 107–145, 2001.

HARTIGAN, J. A.; WONG, M. A. A k-means clustering algorithm. **Applied statistics**, v. 28, p. 100–108, 1979.

HAYKIN, S. **Neural networks and learning machines**. 3rd. ed. Bergen County: Pearson Education, 2009.

HENNIG, C. **fpc: Flexible procedures for clustering**, 2014. R package version 2.1-7. Disponível em: <http://CRAN.R-project.org/package=fpc>.

HOLLAND, J. H. **Adaptation in natural and artificial systems**. Ann Arbor: The University of Michigan Press, 1975.

ISHIOKA, T. An expansion of x-means for automatically determining the optimal number of clusters. In: **Proceedings of International Conference on Computational Intelligence**, 2005. p. 91–96.

KOHONEN, T. The self-organizing map. **Proceedings of the IEEE**, v. 78, n. 9, p. 1464–1480, 1990.

KOHONEN, T.; HYNNINEN, J.; KANGAS, J.; LAAKSONEN, J. **SOM-PAK: The self-organizing map program package**, 1996. Disponível em: <http://www.cis.hut.fi/research/papers/som_tr96.ps.Z>.

LANGLEY, P.; SIMON, H. A. Applications of machine learning and rule induction. **Communications of the ACM**, v. 38, n. 11, p. 54–64, 1995.

LICHMAN, M. **UCI Machine Learning Repository**. 2013. Disponível em: <http://archive.ics.uci.edu/ml>.

LINOFF, G. S.; BERRY, M. J. A. **Data mining techniques: for marketing, sales, and customer relationship management**. 3rd. ed. Danvers: John Wiley & Sons, 2011.

LIU, Y.; LI, Z.; XIONG, H.; GAO, X.; WU, J. Understanding of internal clustering validation measures. In: **Proceedings of the 10th IEEE International Conference on Data Mining (ICDM 2010)**, 2010. p. 911–916.

LIU, Y.; LI, Z.; XIONG, H.; GAO, X.; WU, J.; WU, S. Understanding and enhancement of internal clustering validation measures. **IEEE Transactions on Cybernetics**, v. 43, n. 3, p. 982–994, 2013.

LLOYD, S. Least squares quantization in pcm. **IEEE Transactions on Information Theory**, v. 28, n. 2, p. 129–137, 1982.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**, 1967. v. 1, n. 14, p. 281–297.

MARSLAND, S. **Machine learning: an algorithmic perspective**. Boca Raton: CRC Press, 2009.

MAULIK, U.; BANDYOPADHYAY, S. Performance evaluation of some clustering algorithms and validity indices. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 24, n. 12, p. 1650–1654, 2002.

MOUNT, D. W. **Bioinformatics: sequence and genome analysis**. 2nd. ed. Cold Spring Harbor: CSH Press, 2004.

MUFTI, G. B.; BERTRAND, P.; MOUBARKI, E. Determining the number of groups from measures of cluster stability. In: **Proceedings of the 11th International Symposium on Applied Stochastic Models and Data Analysis**, 2005. p. 17–20.

MULDER, W. D. Instability and cluster stability variance for real clusterings. **Information Sciences**, v. 260, p. 51–63, 2014.

MULDER, W. D.; KUIPER, M.; BOEL, R. Clustering of gene expression profiles: creating initialization-independent clusterings by eliminating unstable genes. **Journal of integrative bioinformatics**, v. 7, n. 3, p. 134, 2010.

NHGRI. **Genome-wide Association Studies**, 2015. Disponível em: <http://www.genome.gov/20019523>.

OLSON, D. L.; DELEN, D. **Advanced data mining techniques**. Heidelberg: Springer-Verlag, 2008.

R Core Team. **R: A Language and Environment for Statistical Computing**, 2015. Disponível em: <http://www.R-project.org/>.

RUTKOWSKI, L. **Computational intelligence: methods and techniques**. Heidelberg: Springer-Verlag, 2008.

SILVA, J. R. **Sistemas de detecção de intrusão com técnicas de inteligência artificial**. Dissertação (Mestrado) — Departamento de Informática, Universidade Federal de Viçosa, Viçosa, 2011.

SINGH, Y.; BHATIA, P. K.; SANGWAN, O. A review of studies on machine learning techniques. **International Journal of Computer Science and Security**, v. 1, n. 1, p. 70–84, 2007.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. Boston: Addison-Wesley, 2006.

WILLIGHAGEN, E. **genalg: R Based Genetic Algorithm**, 2014. R package version 0.1.1.1. Disponível em: <http://CRAN.R-project.org/package=genalg>.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical machine learning tools and techniques**. Burlington: Morgan Kaufmann, 2011.

YAN, J. **som: Self-Organizing Map**, 2010. R package version 0.3-5. Disponível em: <http://CRAN.R-project.org/package=som>.

Apêndice A - DISTRIBUIÇÃO DAS AMOSTRAS INSTÁVEIS QUE PRESERVAM A ESTRUTURA DA BASE DE DADOS DE SNPS REMOVIDAS POR CADA ABORDAGEM

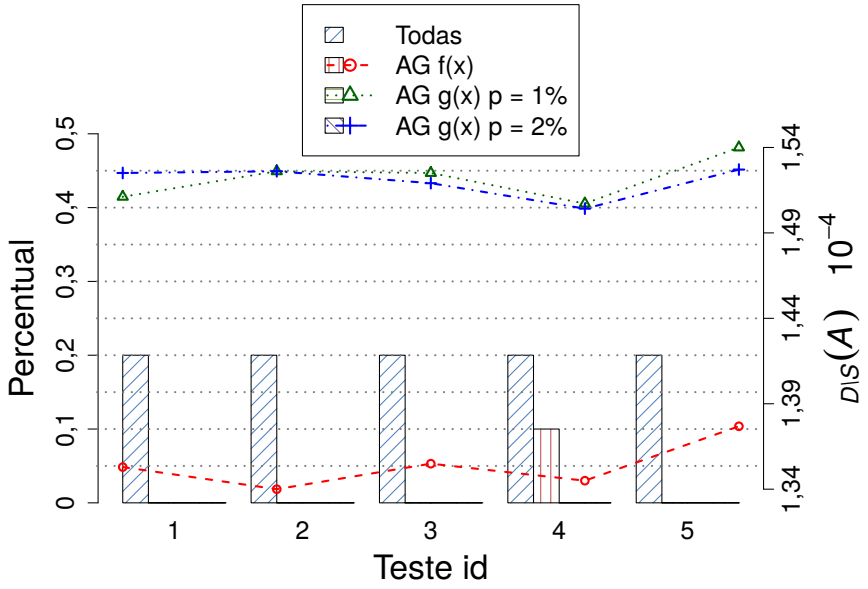

Figura A.1: Distribuição das amostras instáveis que preservam a estrutura dos dados dentro da raça Jersey da base de dados de SNPs removidas utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a primeira e a segunda função de avaliação (Equações 3.3 e 3.4, $f(x)$ e $g(x)$, respectivamente). A instabilidade original (sem remoção de amostras) é $7{,}28 \times 10^{-4}$ e a instabilidade resultante após a remoção de todas as amostras é $19{,}81 \times 10^{-4}$.

Apêndice B - DISTRIBUIÇÃO DAS AMOSTRAS INSTÁVEIS QUE PRESERVAM A ESTRUTURA DA BASE DE DADOS WINE REMOVIDAS POR CADA ABORDAGEM

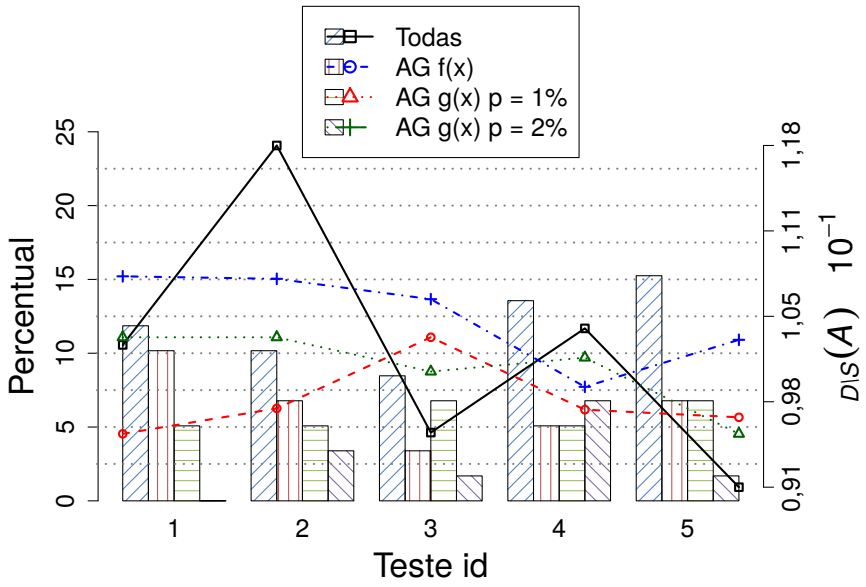

Figura B.1: Distribuição das amostras instáveis que preservam a estrutura dos dados dentro do grupo *tipo 1* da base de dados *Wine* removidas utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a primeira e a segunda função de avaliação (Equações 3.3 e 3.4, $f(x)$ e $g(x)$, respectivamente). A instabilidade original (sem remoção de amostras) é $1{,}38 \times 10^{-1}$.

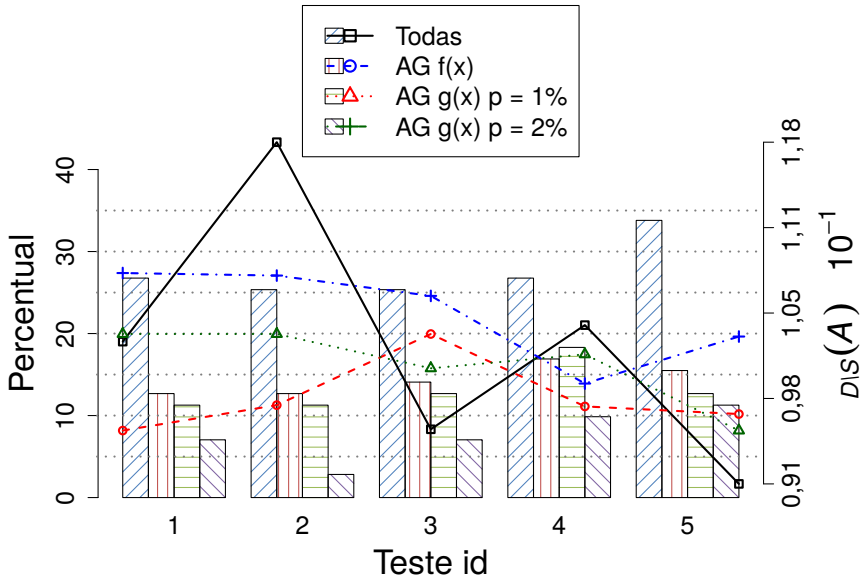

Figura B.2: Distribuição das amostras instáveis que preservam a estrutura dos dados dentro do grupo *tipo 2* da base de dados *Wine* removidas utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a primeira e a segunda função de avaliação (Equações 3.3 e 3.4, $f(x)$ e $g(x)$, respectivamente). A instabilidade original (sem remoção de amostras) é $1{,}38 \times 10^{-1}$.

Apêndice C - DISTRIBUIÇÃO DAS AMOSTRAS INSTÁVEIS QUE PRESERVAM A ESTRUTURA DA BASE DE DADOS IRIS REMOVIDAS POR CADA ABORDAGEM

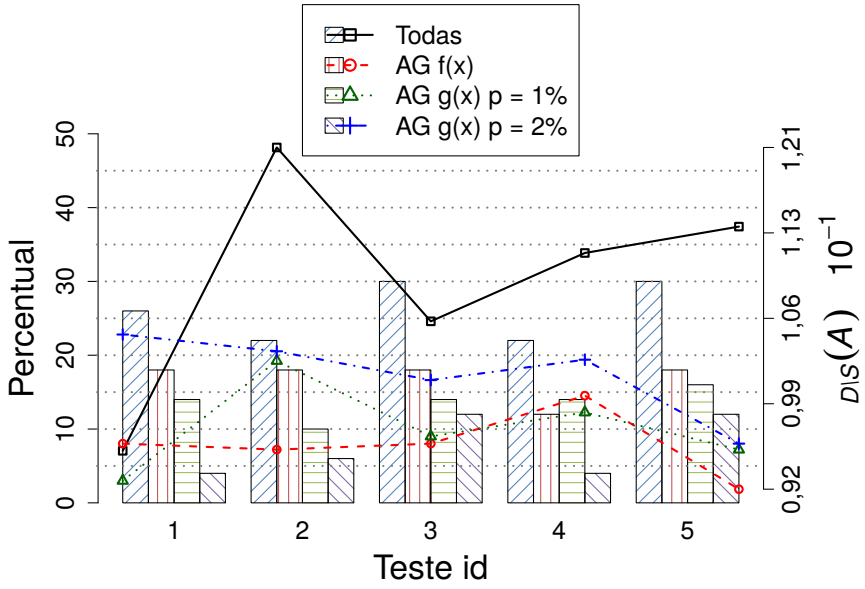

Figura C.1: Distribuição das amostras instáveis que preservam a estrutura dos dados dentro do grupo *Iris versicolor* da base de dados *Iris* removidas utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a primeira e a segunda função de avaliação (Equações 3.3 e 3.4, $f(x)$ e $g(x)$, respectivamente). A instabilidade original (sem remoção de amostras) é $1{,}24 \times 10^{-1}$.

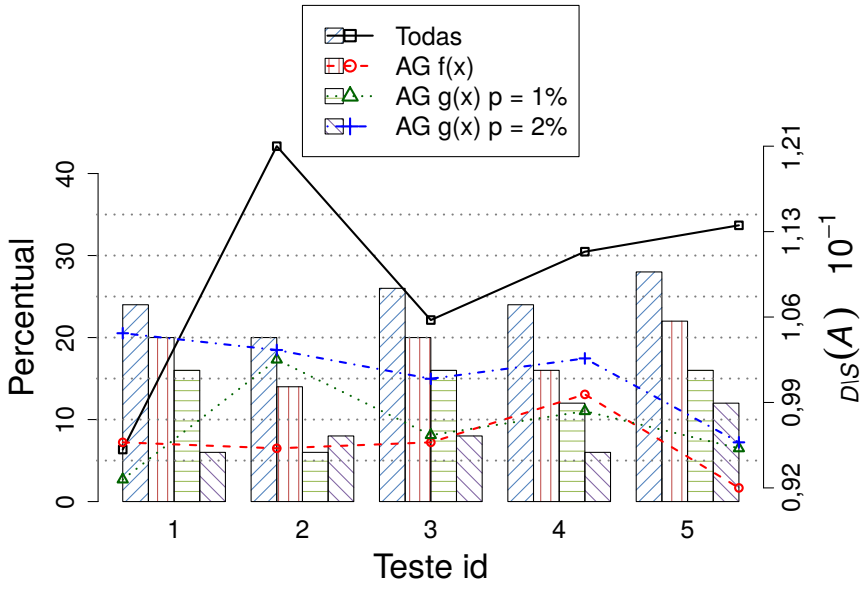

Figura C.2: Distribuição das amostras instáveis que preservam a estrutura dos dados dentro do grupo *Iris virginica* da base de dados *Iris* removidas utilizando o critério de remoção de todas as amostras, sugerido por Mulder (2014), e o critério de seleção e remoção sugerido neste trabalho (Seção 3.5) utilizando a primeira e a segunda função de avaliação (Equações 3.3 e 3.4, $f(x)$ e $g(x)$, respectivamente). A instabilidade original (sem remoção de amostras) é $1{,}24 \times 10^{-1}$.