

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO RIO GRANDE
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

**UM ALGORITMO DE AGRUPAMENTO DE DADOS
BASEADO EM CONSISTÊNCIA COGNITIVA**

por

Thales Vaz Maciel

Tese Submetida ao
PPG em Modelagem Computacional

Rio Grande, Abril, 2022

Thales Vaz Maciel

“Um Algoritmo de Agrupamento de Dados Baseado em Consistência Cognitiva”

Tese apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal do Rio Grande - FURG, como requisito parcial para obtenção do Grau de Doutor. Área de concentração: Modelagem Computacional.

Aprovada em:

BANCA EXAMINADORA

Documento assinado digitalmente
 Leonardo Ramos Emmendorfer
Data: 03/05/2022 18:07:14-0300
Verifique em <https://verificador.iti.br>

Prof. Dr. Leonardo Ramos Emmendorfer (Orientador – FURG)

Documento assinado digitalmente
 Diana Francisca Adamatti
Data: 09/05/2022 08:31:51-0300
Verifique em <https://verificador.iti.br>

Profa. Dra. Diana Francisca Adamatti (FURG)

Documento assinado digitalmente
 PAUL GERHARD KINAS
Data: 09/05/2022 07:47:13-0300
Verifique em <https://verificador.iti.br>

Prof. Dr. Paul Gerhard Kinas (FURG)

Documento assinado digitalmente
 SANDRO DA SILVA CAMARGO
Data: 04/05/2022 18:59:07-0300
Verifique em <https://verificador.iti.br>

Prof. Dr. Sandro da Silva Camargo (Unipampa)

Ficha Catalográfica

M152a Maciel, Thales Vaz.

Um algoritmo de agrupamento de dados baseado em consistência cognitiva / Thales Vaz Maciel. – 2022.

100 f.

Tese (doutorado) – Universidade Federal do Rio Grande – FURG, Programa de Pós-Graduação em Modelagem Computacional, Rio Grande/RS, 2022.

Orientador: Dr. Leonardo Ramos Emmendorfer.

1. Agrupamento de dados 2. Dissonância cognitiva 3. Equilíbrio psicológico I. Emmendorfer, Leonardo Ramos II. Título.

CDU 004.021

Catálogo na Fonte: Bibliotecário José Paulo dos Santos CRB 10/2344

AGRADECIMENTOS

Agradeço ao meu orientador, Prof. Dr. Leonardo Ramos Emmendorfer, pelos ensinamentos, aconselhamentos, reuniões, revisões de trabalhos, paciência, dedicação e parceria. Sua orientação foi fundamental para que este trabalho de doutorado e outros relacionados se concretizassem.

Agradeço aos colegas e demais professores do PPGMC, que assistiram minhas apresentações e seminários e contribuíram para o melhoramento do meu trabalho.

Agradeço à Universidade Federal do Rio Grande, que me acolheu desde o mestrado no PPGComp e, agora, no doutorado no PPGMC.

Agradeço à minha família por todo o suporte e amor que me dão. Dedico este trabalho à vocês.

E agradeço à Deus, pelas bênçãos, oportunidades e caminhos.

RESUMO

O agrupamento de dados trata da descoberta de agrupamentos de dados por meio de aprendizado de máquina, de forma que os elementos integrantes de cada grupo sejam o mais semelhantes entre si o quanto possível e os integrantes de grupos diferentes sejam o mais dissimilares entre si o quanto possível. A literatura apresenta diversos algoritmos sobre esta temática, inclusive abrangendo a modelagem do comportamento dos indivíduos em sociedades de diferentes espécies. A inovação do presente trabalho está em que o algoritmo de agrupamento de dados Cognitive Consistency Clustering (C3) é fundamentado na teoria social da dissonância cognitiva e na teoria do balanço psicológico. O trabalho investiga, ainda, a influência da capacidade de tomada de decisão pautada pelo teor psicológico das opções na tarefa de agrupamento de dados em ambiente computacional. Os resultados obtidos mostraram superioridade de desempenho do algoritmo proposto, com ganhos médios de 10,4% e 7,7%, na comparação com os algoritmos K-means e AVC, respectivamente, sob óptica de métricas de qualidade internas.

Palavras-chaves: agrupamento de dados, dissonância cognitiva, equilíbrio psicológico

ABSTRACT

Data clustering is about discovering partitions in the data by means of machine learning whereas items from the same groups are as similar to each other as possible and items from different groups are as dissimilar to each other as possible. Literature presents vast diversity on data clustering approaches, including systems that model the behavior of individuals in societies of different species. The innovation of the present work is in that the Cognitive Consistency Clustering (C3) algorithm is reasoned upon the social theory of cognitive dissonance and the psychological theory of balance. This work also investigates the psychological balance aware decision making capability's influence on the data clustering task in a computational environment. Obtained results show superiority of the proposed algorithm (C3) over K-means and AVC, with average quality improvements of 10,4% and 7,7% respectively, according to internet cluster quality metrics.

Palavras-chaves: data clustering, cognitive dissonance, balance theory

ÍNDICE

1	Introdução	12
1.1	Justificativa	12
1.2	Objetivos	12
1.3	Estrutura do Trabalho	13
2	Conceitos Preliminares	14
2.1	Agrupamento de Dados (<i>Clustering</i>)	14
2.2	Redução de Dimensionalidade	16
2.3	Avaliação de Qualidade de Agrupamentos (<i>Clusters</i>)	17
2.4	Análise de Variância e Teste de Tukey	22
2.5	Dissonância Cognitiva e Balanceamento Psicológico	24
3	Revisão Bibliográfica	28
3.1	Trabalhos Relacionados	28
3.2	Síntese da Revisão	46
4	Metodologia	48
4.1	Pré-processamento e Inicialização dos Agentes	48
4.2	Posicionamento e Movimentação dos Agentes	49
4.3	Tomada de Decisão do Próximo Movimento pelos Agentes	50
5	Resultados Obtidos	58
5.1	Wine Dataset	59
5.2	Zoo Dataset	65
5.3	Thyroid Dataset	71
5.4	FLAME Dataset	77
5.5	<i>Escherichia Coli</i> Dataset	83
6	Discussão	90
7	Conclusão	92
8	Referências	94

LISTA DE FIGURAS

2.1	Exemplos de tríades de relacionamento positivo e negativo, adaptado de Cartwright e Harary (1956) [10].	26
3.1	Protocolo de revisão realizado.	29
3.2	Taxonomia de algoritmos de agrupamento de dados.	30
4.1	Distribuição de agentes no ambiente 2D.	49
4.2	Círculo de percepção de agente em destaque.	49
4.3	Círculo de percepção do agente em destaque dividido em seções.	50
4.4	Grafo de relações positivas encontradas no ambiente.	51
4.5	Detecção do teor psicológico por setores do círculo de percepção do agente em destaque.	52
4.6	Exemplo de situação em que o agente em destaque não encontra setor do seu círculo de percepção com teor psicológico positivo, o que o impede de se movimentar na respectiva iteração.	54
4.7	Conformação final dos agentes do sistema.	55
4.8	Tradução da conformação final dos agentes do sistema em agrupamentos de dados.	57
5.1	Comparativo visual de desempenho dos algoritmos K-means, AVC e C3 (Wine).	64
5.2		65
5.3	Comparativo visual de desempenho dos algoritmos K-means, AVC e C3 (Zoo).	70
5.4		71
5.5	Comparativo visual de desempenho dos algoritmos K-means, AVC e C3 (Thyroid).	76
5.6		77
5.7	Comparativo visual de desempenho dos algoritmos K-means, AVC e C3 (FLAME).	82
5.8		83
5.9	Comparativo visual de desempenho dos algoritmos K-means, AVC e C3 (<i>Escherichia Coli</i>).	88
5.10		89
6.1	Comparativo de Diferenças de Desempenho Médio do algoritmo C3 sobre os algoritmos K-means e AVC.	91

LISTA DE TABELAS

1	Parâmetros utilizados nos testes sobre o conjunto de dados de vinhos. . .	59
2	Indicadores de qualidade de agrupamentos de dados (Wine).	64
3	Parâmetros utilizados nos testes sobre o conjunto de dados de animais do zoológico.	66
4	Indicadores de qualidade de agrupamentos de dados (Zoo).	70
5	Parâmetros utilizados nos testes sobre o conjunto de dados thyroid. . . .	72
6	Indicadores de qualidade de agrupamentos de dados (Thyroid).	76
7	Parâmetros utilizados nos testes sobre o conjunto de dados FLAME. . . .	78
8	Indicadores de qualidade de agrupamentos de dados (FLAME).	82
9	Parâmetros utilizados nos testes sobre o conjunto de dados de Ecoli. . . .	84
10	Indicadores de qualidade de agrupamentos de dados (<i>Escherichia Coli</i>). .	88

LISTA DE SÍMBOLOS

$\leftarrow$	É definido como.
$\in$	Pertence ao conjunto.
P	Uma pessoa, no contexto do modelo POX.
O	Outra pessoa, no contexto do modelo POX.
X	Um objeto de interesse, no contexto do modelo POX.
L	Uma relação positiva, no contexto do modelo POX.
L'	Uma relação negativa, no contexto do modelo POX.
U	Uma relação de unidade, no contexto do modelo POX.

Símbolos gregos

Δ	Conjunto de dados.
δ	Ponto de dado no conjunto de dados.
Γ	Raio de percepção dos agentes do sistema.
λ	Extensão do passo de movimentação dos agentes no sistema.
ϕ	Quantidade de setores a dividir o círculo de percepção dos agentes no sistema.
ϵ	Limiar de satisfação cognitiva.
ϖ	Quantidade de relações positivas de um agente.
A	Conjunto de agentes do sistema.
α	Agente do sistema.
Ψ	Conjunto de relações entre os agentes do sistema.
ψ	Uma relação entre dois agentes do sistema.
Ω	Conjunto de setores de um círculo de percepção de um agente.
ω	Um setor do círculo de percepção de um agente.
τ	Um setor do círculo de percepção de um agente que foi escolhido para pautar sua próxima movimentação.
κ	Um rótulo de agrupamento de dados.
P	Subconjunto de relações entre os agentes do sistema.
ρ	Um relação entre dois agentes do sistema, similar a ψ .
N	Conjunto de tríades de relações interagentes.
ν	Uma tríade de relações interagentes.
M	Conjunto de produtos de tríades de relações interagentes.
μ	Um produto de tríade de relações interagentes.
ι	Limiar de satisfação cognitiva dos agentes do sistema.
Z	Grau de balanceamento de um conjunto de tríades de relações interagentes.
ζ	Teor psicológico de um conjunto de tríades de relações interagentes.

LISTA DE ABREVIATURAS

ACOCC	Ant Colony Optimization Clustering with Chaos
ANOVA	Analysis of Variance
AVC	Agent View Clustering
BPEC	Belief Peaks Evidential Clustering
C3	Cognitive Consistency Clustering
CF+-ERC	Tree-Based Effective Multiple Range Queries-Based Clustering
CMNC	Closeness Measure based on Neighborhood Chain
CRS	Clustering based on Region Segmentation
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
dFC	Data-Density-Aware Fuzzy Clustering
DPC-SNR	Density Peaks Clustering using Symmetric Neighborhood Relationship
FensiVAT	Fast clustering by combining ENsembles of random projections with Scalable version of Improved Visual Assessment of Clustering Tendency
FINCH	First-Integer-Neighbor-based Clustering Hierarchy
FIPS	Fully-Informed Particle Swarm
FLAME	Fuzzy clustering by Local Approximation of MEMbership
FPS	Fitness Proportionate Share
HPFCM	Hybrid and Parameter-free Clustering Method
HSD	Honestly Significant Difference
IWGC	Internally Weighted Graph Cut
LLK-HKAP	Local Linear Embedding Hybrid Kernel Affinity Propagation
MulSim	Similar-to-Multiple-Point Clustering
NIPC	Novel Insect Pollination Clusterin
PCA	Principal Component Analysis
PSO	Particle Swarm Optimization
OLO	Optimal Leaf Optimization
RAM	Recursive Agent-Based Mixing
RDC	Robust Dual Clustering with Adaptive Manifold Regularization
RDC-GP	Robust Density Clustering Algorithm Based on Gravity Peak
SCA	Swarm Clustering Algorithm
SDE	Sparse Density Entropy
TGMM	Transfer Clustering based on Gaussian Mixture Model
USC	Uncertain Spectral Clustering

1 INTRODUÇÃO

O agrupamento de dados, ou *clustering*, pode ser definido como a descoberta de agrupamentos em conjuntos de dados de forma automatizada em ambiente computacional, por meio de aprendizado de máquina. O objetivo deste tipo de tarefa é que os agrupamentos resultantes contenham elementos tão similares entre os que compartilham o mesmo agrupamento o quanto possível e simultaneamente tão dissimilares com aqueles pertencentes a outros agrupamentos o quanto possível.

Ocorre que, desde a publicação do algoritmo de Lloyd em 1982 [44], muitos algoritmos cujo objetivo compartilha desta temática foram publicados. Desta forma, é notória a pluralidade de estratégias para a solução do problema, desde a utilização de pseudo-centros que se adaptam recorrentemente até encontrar estabilidade [44, 28], a detecção de declínios de densidade de dados que se traduzem em divisórias entre agrupamentos [77, 80] e modelagem de processos biológicos, em que o comportamento natural conhecido de alguma espécie apresente algum viés na produção de agrupamentos de indivíduos [84, 52].

Também é notória a diversidade de aplicações do agrupamento de dados, desde o processamento de imagens para compactação de mídias, descoberta de perfis e recomendação de consumo de mercado, análise de redes sociais e detecção de outliers.

1.1 Justificativa

Embora tenham sido produzidos algoritmos de agrupamento de dados inspirados no comportamento social dos indivíduos de diversas espécies, incluindo citações ao comportamento em face de cognições conflitantes, por exemplo, eles não modelam aspectos específicos dos processos psicológicos que norteiam o comportamento dos indivíduos que compõem uma sociedade.

Neste contexto, o presente trabalho aprecia as teorias da dissonância cognitiva, conforme Festinger (1957) [20], e do equilíbrio psicológico, conforme Heider (1958) [26] e Cartwright e Harary (1956) [10]. Visa-se a hipótese de que a aplicação de tais teorias nos processos de tomada de decisão em tarefas de aprendizado de máquina não supervisionado traz ganhos de desempenho aos modelos descobertos, na comparação com abordagens tradicionais.

Em outras palavras, espera-se que um algoritmo de agrupamento de dados, dotado da capacidade de interpretação de aspectos cognitivos dos indivíduos presentes em um sistema, seja capaz de detectar relações interindividuais que, em abordagens tradicionais, são desconhecidas e, portanto inexploradas, por fim apresentando melhoria de desempenho em cenários específicos na comparação com algoritmos relacionados.

1.2 Objetivos

O objetivo do presente trabalho é definir como a modelagem e aplicação de modelos de processos cognitivos humanos, referentes ao equilíbrio psicológico dos indivíduos, em tarefas de

agrupamento de dados pode influenciar na qualidade de modelos descobertos. Especificamente, objetiva-se:

- a definição de um algoritmo original para agrupamento de dados, apropriado das teorias do balanceamento psicológico e da dissonância cognitiva; e
- a eventual superioridade na comparação com métodos relacionados, presentes na bibliografia, no tocante a qualidade dos modelos descobertos e acurácia na definição dos agrupamentos.

1.3 Estrutura do Trabalho

Conceitos fundamentais acerca do agrupamento de dados, redução de dimensionalidade e métricas de qualidade de agrupamentos de dados são apresentados no capítulo 2. Este capítulo também sintetiza as teorias da dissonância cognitiva e do balanceamento psicológico, as responsáveis por nortear o comportamento do sistema de agrupamento de dados desenvolvido no presente trabalho.

O capítulo 3 apresenta uma revisão de trabalhos correlatos, especificamente algoritmos de agrupamento de dados, que foram publicados recentemente. No total, 26 abordagens são descritas, posicionando o presente trabalho em meio a bibliografia, de forma a evidenciar o ineditismo da presente proposta.

O capítulo 4 detalha as etapas de execução do algoritmo apresentado neste trabalho, especificando os mecanismos de tomada de decisão realizadas ao longo do processo de agrupamento de dados. São destacadas as aplicações dos conceitos inerentes às teorias da dissonância cognitiva e balanço psicológico e como elas influenciam o comportamento dos multiagentes de um sistema.

Os resultados obtidos são apresentados no capítulo 5, na forma comparação de desempenho entre o algoritmo proposto, denominado Cognitive Consistency Clustering (C3), e os algoritmos K-means e Agent View Clustering (AVC). Esta comparação foi feita pela perspectiva de 9 métricas de qualidade de agrupamento de dados, 6 externas e 3 internas, sobre 5 conjuntos de dados distintos. É evidenciada a influência das teorias aplicadas na qualidade dos modelos descobertos pela abordagem proposta, através da comparação com modelos descobertos por algoritmos que não são dotados de tais mecanismos.

Os resultados apresentados no capítulo 5 são discutidos no capítulo 6, onde também são apresentadas estatísticas genéricas sobre os resultados dos experimentos. O trabalho é concluído no capítulo 7, onde os objetivos são revisitados com comentários sobre os desempenhos perante os diversos tipos de métricas de qualidade de agrupamentos utilizadas no estudo.

2 CONCEITOS PRELIMINARES

Este capítulo apresenta dois temas de interesse ao presente trabalho. Trata-se do agrupamento de dados, como tarefa de descoberta de conhecimento da mineração de dados e que compreende o contexto onde o presente trabalho está inserido, e as teorias da dissonância cognitiva e balanceamento psicológico, da psicologia, que explicam o comportamento humano em situações de tomada de decisão frente cognições dissonantes e a análise do teor psicológico de situações alternativas.

2.1 Agrupamento de Dados (*Clustering*)

Agrupamento de dados, também apresentado como *clustering*, refere-se a uma tarefa de mineração de dados cujo objetivo é encontrar agrupamentos em um conjunto de dados, de forma automática e não supervisionada e de forma que os elementos que pertencem aos mesmos agrupamentos sejam o mais semelhantes entre si o quanto possível e o mais dissimilar em relação aos elementos pertencentes aos agrupamentos diferentes do seu [43, 74, 51].

Algoritmos de agrupamento de dados podem ser categorizados de acordo com o mecanismo ou estrutura no qual baseiam o aprendizado: hierarquia, computação difusa, pseudo-centros, busca, grafos, malhas, densidade, modelos e subespaços [28].

O agrupamento de dados baseado em hierarquia inclui as tipologias aglomerativas e divisivas. Os algoritmos aglomerativos inicialmente definem que cada ponto de dado inicial compreende um agrupamento distinto e os mescla iterativamente de acordo com algum critério definido. Os algoritmos divisivos, por sua vez, inicialmente definem um único agrupamento, ao qual pertencem todos os pontos de dados de entrada, iterativamente dividindo os agrupamentos, como dendogramas, de acordo com algum critério definido. De acordo com Li et al. (2018) [41], este tipo de algoritmo de agrupamento de dados geralmente é sensível a ordem em que os pontos de dados são processados e sofrem de problemas relacionados à alta complexidade computacional. Exemplos são os algoritmos Cure [31], Chameleon [37] e Rock [32].

O agrupamento de dados tradicional sugere que cada ponto de dado pertence a um único agrupamento [28]. O agrupamento de dados difuso aplica conceitos da computação difusa de forma a conceber que cada ponto de dado pertence a todos os agrupamentos de acordo com uma função de pertencimento. Exemplos de algoritmos de agrupamento de dados que utilizam da computação difusa são o Fuzzy K-means [61] e o Fuzzy C-means [18, 6].

No método de agrupamento de dados baseado em pseudo-centros, a quantidade de agrupamentos a serem descobertos deve ser definida preliminarmente, mesmo que por informação do usuário. Assim, são criados pseudo-centros para cada agrupamento a ser descoberto, cujas posições no espaço de dados são definidas aleatoriamente ou de acordo com alguma heurística.

A partir de então, cada ponto de dado é atribuído ao pseudo-centro mais próximo de si e o posicionamento do pseudo-centro redefinido de acordo com algum critério, recorrentemente, até que não haja alteração no estado dos agrupamentos em uma iteração completa [28]. Contudo,

os agrupamentos descobertos por este método são vulneráveis ao posicionamento inicial dos pontos de dados e são passíveis de incorreção quando forem de diferentes formatos e tamanhos [41]. Exemplos deste tipo de algoritmo são os algoritmos K-means [44], K-medoids [38] e K-medians [36].

Algoritmos de agrupamento de dados baseado em busca podem ser definidos como uma extensão de quaisquer outras tipologias, visto que eles são projetados de forma a evitar um retorno de agrupamentos baseados em soluções ótimas locais [28]. Eles propositalmente exploram o espaço de dados além da otimização local de forma a encontrar agrupamentos que caracterizem o ótimo global, o que pode ser implementado como extensão em diversos algoritmos. Exemplos deste tipo de algoritmo são os algoritmos Ant Colony Optimization-Based Clustering Algorithm with Chaos (ACOCC) [79] e Swarm Clustering Algorithm (SCA) [84].

Algoritmos de agrupamento de dados baseados em grafos inicialmente constroem um grafo a partir do conjunto de pontos de dado de entrada e, então, aplicam um algoritmo de agrupamento de dados adequado ao particionamento do grafo. Estas partições são, então, traduzidas em agrupamentos de dados. Nesta categoria, são considerados aqueles algoritmos baseados em encadeamento, pois as ligações entre pontos de dados podem ser consideradas arestas entre nós de um grafo.

Desta forma, este tipo de abordagem pode ser considerado uma extensão de outras, visto que apenas traz uma óptica diferenciada ao conjunto de dados analisado, a exemplo das abordagens baseadas em busca. Exemplos de algoritmos que utilizam esta abordagem são o Optimal Leaf Optimization (OLO) [54], Closeness Measure based on Neighborhood Chain (CMNC) [42] e Density Peaks Clustering using Symmetric Neighborhood Relationship (DPC-SNR) [77].

Métodos de agrupamento de dados baseados em malhas abstraem o espaço ocupado pelos dados em um conjunto finito de células que constituem uma estrutura de dado em formato de malha. Em outras palavras, difere-se de outras abordagens ao agrupamento de dados no sentido de que não trata os pontos de dados diretamente, mas, sim, o espaço que compreende a distribuição dos pontos de dados no plano euclidiano.

Este tipo de algoritmo divide o processamento em cinco etapas [30]: criação da estrutura de malha, com o particionamento do espaço de dados em quantidade finita de células, o cálculo da densidade correspondente a cada célula, a ordenação das células em função de suas densidades, identificação dos centros dos agrupamentos de dados e a fusão das células vizinhas com aquelas identificadas como centros de agrupamentos.

Esta abordagem ao agrupamento de dados trata um conflito entre a eficiência do algoritmo e a qualidade dos agrupamentos descobertos [41]. Exemplos de algoritmos que utilizam este mecanismo são o Statistical Information Grid Clustering Algorithm [73] e Grid Clustering [60].

O agrupamento de dados baseado em densidade é capaz de encontrar agrupamentos de formatos arbitrários, visto que os agrupamentos são definidos como regiões de alta densidade separadas por regiões de baixa densidade. Segundo Gan, Ma e We (2007) [28], este tipo de algoritmo necessita de apenas uma leitura do conjunto de dados, o que reduz a complexidade computacional em comparação com outras abordagens, é capaz de tratar dados ruidosos e

dispensa a informação da quantidade de agrupamentos a serem descobertos, o que é definido automaticamente conforme as diferenças de densidade no espaço de dados são detectadas. Exemplos de algoritmos que utilizam esta abordagem são o Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [19], Data-Density-Aware Fuzzy Clustering (dFC) [71] e Sparse Density Entropy (SDE) [41].

Algoritmos de agrupamento de dados baseado em modelos de probabilidade utilizam modelos paramétricos para otimizar a aptidão dos modelos sobre o conjunto de dados. Estes modelos são referentes aos tipos de restrições e propriedades geométricas das matrizes de covariância descobertas.

Neste tipo de abordagem, é conceitualizado que os dados são oriundos de uma mistura de distribuições de probabilidade, de forma que cada mistura representa um agrupamento diferente. Portanto, métodos específicos têm melhor desempenho conforme os dados mais se adequam ao modelo utilizado nos agrupamentos. Exemplos de algoritmos que utilizam este tipo de abordagem são o Expectation Maximization [16] e Transfer Clustering based on Gaussian Mixture Model (TGMM) [72].

O agrupamento de dados por sub-espacos é uma forma de abordar o problema quando o conjunto de dados analisado tem alta dimensionalidade. A técnica encontra agrupamentos de dados em diferentes seleções de uma ou mais dimensões do conjunto de dados, o que caracteriza os sub-espacos.

Este método de agrupamento de dados pode ser diferenciado em duas tipologias, sendo de baixo à cima, onde o algoritmo inicia pela descoberta de agrupamentos em uma dimensão e iterativamente os une para processar espacos de maiores dimensões, ou de cima à baixo, onde agrupamentos são encontrados no conjunto total das dimensões e avalia cada agrupamento em termos das dimensões que os compreendem. Exemplos deste tipo de algoritmo são o Clustering in Quest [2] e o Fast and INtelligent subspace clustering algorithm using DIMension voTING [76].

2.2 Redução de Dimensionalidade

Sob a perspectiva de bancos de dados, redução de dimensionalidade significa o mapeamento de um conjunto de dados de alta dimensionalidade, em seu estado original, para uma representação de menor dimensionalidade que seja capaz de capturar o conteúdo original do conjunto de dados de acordo com algum critério [63].

Conjuntos de dados de alta dimensionalidade são frequentemente estudados em mineração de dados e estatística. A redução de dimensionalidade é apresentada como solução para eliminação de dimensões ruidosas, o que impacta positivamente na qualidade de modelos descobertos por tarefas de classificação e agrupamento de dados, por exemplo, além da redução do custo computacional de execução [17].

O desafio maior da redução de dimensionalidade é a manutenção da qualidade do conjunto de dados resultante, em comparação com o conjunto de dados original. Métodos conhecidos

para tal incluem a transformada discreta de Fourier [75], a transformada discreta de Wavelet [64], a aproximação constante adaptável por partes [39], a projeção de preservação de localização [33] e a análise de componentes principais (PCA) [50], sendo a última a fundamentação para a criação de técnicas de redução de dimensionalidade modernas e notavelmente a mais adotada em estudos que envolvem este tipo de tarefa [17].

No contexto do presente trabalho, houveram diversas aplicações do método PCA para redução de dimensionalidade dos conjuntos de dados analisados. O motivo por sua opção foi a linearidade inerente ao método [50] que, conseqüentemente, mantém a ocasional indistinção linear entre os agrupamentos do conjunto de dados, quando da redução de dimensionalidade no mesmo. Ocorre que o algoritmo de agrupamento de dados proposto tem sua maior contribuição nestes cenários, descobrindo relações entre os pontos de dados que seriam despercebidas em abordagens relacionadas¹.

2.3 Avaliação de Qualidade de Agrupamentos (*Clusters*)

O índice de Rand é uma métrica de similaridade entre dois modelos de agrupamento de dados [53]. Se relaciona ao conceito de acurácia em que verifica a classificação dos elementos conforme realizada de forma não supervisionada na comparação com as classes verdadeiras dos elementos analisados. O índice da Rand ajustado é um melhoramento do índice original pela robustez perante permutações e normalizações ao acaso [35]. O domínio da métrica compreende o espectro de -1 a 1, onde -1 identifica classificações totalmente independentes e 1 identifica classificações totalmente consonantes.

A formulação matemática para o índice de Rand é apresentada na equação 2.1, onde:

- a representa a quantidade de pares de elementos que estão nos mesmos agrupamentos originais e nos mesmos agrupamentos no modelo descoberto;
- b representa a quantidade de pares de elementos que estão em agrupamentos originais diferentes e em agrupamentos diferentes no modelo descoberto;
- C_2^n representa o total de possíveis pares de elementos do conjunto de dados.

$$RI = \frac{a + b}{C_2^n} \quad (2.1)$$

O ajuste ao índice de Rand foi proposto [47] na forma de dedução do índice estimado para agrupamentos aleatórios $E[RI]$. Esta formulação é apresentada na equação 2.2.

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \quad (2.2)$$

¹Isto é evidenciado na seção 6, referente à discussão dos resultados obtidos na comparação de desempenho entre a proposta presente e outras relacionadas.

A informação mútua é uma função que mede a concordância entre dois conjuntos de atribuições categóricas [69]. Diferencia-se do índice de Rand em que utiliza conceitos da teoria da informação, como a entropia, que quantifica a informação contida em uma variável aleatória.

A fórmula 2.3 apresenta o cálculo de entropia para um conjunto de agrupamentos, onde:

- u representa um modelo de atribuições de agrupamentos;
- $P(i) = |u_i|/N$ representa a probabilidade de um elemento de u escolhido aleatoriamente pertencer ao agrupamento u_i , em razão da quantidade total de elementos N .

$$H(u) = - \sum_{i=1}^{|u|} P(i) \log(P(i)) \quad (2.3)$$

Da mesma forma, a fórmula 2.4 apresenta o cálculo de entropia para outro conjunto de agrupamentos, onde:

- v representa um modelo de atribuições de agrupamentos (distinto de u);
- $P'(j) = |v_j|/N$ representa a probabilidade de um elemento de v escolhido aleatoriamente pertencer ao agrupamento v_j , em razão da quantidade total de elementos N .

$$H(v) = - \sum_{j=1}^{|v|} P'(j) \log(P'(j)) \quad (2.4)$$

A fórmula 2.5 apresenta equação para o cálculo de informação mútua entre os conjuntos de agrupamentos u e v , onde:

- u e v representam dois modelos de atribuições de agrupamentos distintos;
- $P(i) = |u_i|/N$ representa a probabilidade de um elemento de u escolhido aleatoriamente pertencer ao agrupamento u_i , em razão da quantidade total de elementos N ;
- $P'(j) = |v_j|/N$ representa a probabilidade de um elemento de v escolhido aleatoriamente pertencer ao agrupamento v_j , em razão da quantidade total de elementos N ;
- $P(i, j) = |u_i \cap v_j|/N$ representa a probabilidade de um elemento aleatório do conjunto de dados analisado pertencer às classes u_i e v_j , em razão da quantidade total de elementos N .

$$MI(u, v) = \sum_{i=1}^{|u|} \sum_{j=1}^{|v|} P(i, j) \log \left(\frac{P(i, j)}{P(i)P'(j)} \right) \quad (2.5)$$

A normalização da informação mútua ocorre de forma que valores próximos de 0 indicam categorização aleatória, enquanto valores próximos de 1 indicam a concordância entre os conjuntos de classificações. A equação 2.6 apresenta o cálculo desta normalização, onde:

- $MI(u, v)$ representa a informação mútua entre os dois conjuntos de agrupamentos, u e v ;
- $\overline{\{H(u), H(v)\}}$ representa a média aritmética entre as entropias do conjunto de agrupamentos u e v .

$$NMI(u, v) = \frac{MI(u, v)}{\overline{\{H(u), H(v)\}}} \quad (2.6)$$

Outras métricas para avaliação intuitiva de classes obtidas por aprendizado não supervisionado, que utilizam o conhecimento das classes verdadeiras são a homogeneidade e a completude. A homogeneidade mede o quanto cada agrupamento contém elementos de apenas uma classe, enquanto a completude mede o quanto todos membros de uma classe pertencem ao mesmo agrupamento [56]. Ambas métricas apresentam domínio no espectro de 0 a 1, onde 0 significa o pior valor possível e 1 o melhor valor possível.

A equação 2.7 apresenta a fórmula para cálculo da homogeneidade, onde:

- $H(C|K)$ representa a entropia condicional dos agrupamentos verdadeiros em conhecimento dos agrupamentos descobertos pela tarefa de *clustering*;
- $H(C)$ representa a entropia dos agrupamentos verdadeiros.

$$h = \begin{cases} 1 & \text{se } H(C, K) = 0 \\ 1 - \frac{H(C|K)}{H(C)} & \text{senão} \end{cases} \quad (2.7)$$

A equação 2.8 apresenta a fórmula para cálculo de entropia condicional com aplicação no cálculo de homogeneidade. Esta entropia condicional quantifica a informação necessária para descrever o agrupamento verdadeiro de um elemento do conjunto, sendo que o agrupamento descoberto pelo algoritmo é conhecido. A equação 2.9 apresenta a fórmula para cálculo da entropia dos agrupamentos verdadeiros.

$$H(C|K) = - \sum_{k=1}^{|K|} \sum_{c=1}^{|C|} \frac{a_{ck}}{N} \log \frac{a_{ck}}{\sum_{c=1}^{|C|} a_{ck}} \quad (2.8)$$

$$H(C) = - \sum_{c=1}^{|C|} \frac{\sum_{k=1}^{|K|} a_{ck}}{n} \log \frac{\sum_{k=1}^{|K|} a_{ck}}{n} \quad (2.9)$$

A equação 2.10 apresenta a fórmula para cálculo da completude, onde:

- $H(K|C)$ representa a entropia condicional dos agrupamentos descobertos pela tarefa de *clustering* em conhecimento dos agrupamentos verdadeiros;
- $H(K)$ representa a entropia dos agrupamentos descobertos pela tarefa de *clustering*.

$$c = \begin{cases} 1 & \text{se } H(K, C) = 0 \\ 1 - \frac{H(K|C)}{H(K)} & \text{senão} \end{cases} \quad (2.10)$$

A equação 2.11 apresenta a fórmula para cálculo de entropia condicional com aplicação no cálculo de completude. Esta entropia condicional quantifica a informação necessária para descrever o agrupamento descoberto pelo algoritmo para elemento do conjunto, sendo que o agrupamento verdadeiro é conhecido. A equação 2.12 apresenta a fórmula para cálculo da entropia dos agrupamentos descobertos pela tarefa de *clustering*.

$$H(K|C) = - \sum_{c=1}^{|C|} \sum_{k=1}^{|K|} \frac{a_{ck}}{N} \log \frac{a_{ck}}{\sum_{k=1}^{|K|} a_{ck}} \quad (2.11)$$

$$H(K) = - \sum_{k=1}^{|K|} \frac{\sum_{c=1}^{|C|} a_{ck}}{n} \log \frac{\sum_{c=1}^{|C|} a_{ck}}{n} \quad (2.12)$$

A média harmônica da homogeneidade e completude atribuída a um conjunto de agrupamentos, também conhecida como métrica V, também é utilizada para verificação da qualidade de agrupamentos de dados, também permitindo uma ponderação no peso da homogeneidade na equação [56]. Isto habilita que se adeque o cálculo da métrica conforme a necessidade de casos específicos. A equação 2.13 apresenta a fórmula para cálculo da métrica V.

$$v = \frac{(1 + \beta) \cdot \text{homogeneidade} \cdot \text{completude}}{\beta \cdot \text{homogeneidade} + \text{completude}} \quad (2.13)$$

O índice de Fowlkes-Mallows é definido como a média geométrica entre a precisão e revocação calculados a partir dos pares de classes descobertas pela tarefa de agrupamento de dados e as verdadeiras [25]. Desta forma, o índice é calculado pela razão entre os positivos verdadeiros e a raiz quadrada do produto da soma dos positivos verdadeiros e os falsos verdadeiros e a soma dos positivos verdadeiros e falsos negativos. O domínio dos valores para a métrica é o espectro de 0 a 1, onde quanto mais alto o valor, maior a similaridade entre dois conjuntos de agrupamentos.

A equação 2.14 apresenta a fórmula para cálculo do coeficiente de Fowlkes-Mallows, onde:

- VP representa a quantidade de pares de elementos que pertencem ao mesmo agrupamento em ambos conjuntos de atribuições de agrupamentos, o verdadeiro e o descoberto pelo algoritmo de *clustering*;
- FP representa a quantidade de pares de elementos que pertencem ao mesmo agrupamento no conjunto de atribuições de agrupamentos verdadeiro, mas não no conjunto de agrupamentos descoberto pelo algoritmo de *clustering*;
- FN representa a quantidade de pares de elementos que pertencem ao mesmo agrupamento no conjunto de atribuições de agrupamentos descoberto pelo algoritmo de *clustering*, mas não no conjunto de agrupamentos verdadeiro.

$$FM = \frac{VP}{\sqrt{(VP + FP) * (VP + FN)}} \quad (2.14)$$

O coeficiente de Silhouette verifica a qualidade de um modelo de agrupamento de dados a partir do próprio modelo, sem necessidade de verificação das classes verdadeiras de cada elemento [57]. Para cada instância, o coeficiente de Silhouette é calculado como a razão entre a subtração da média de distância entre a instância em questão e todas outras instâncias da mesma classe da média de distância entre a instância em questão e todas as outras instâncias no agrupamento mais próximo do qual ela pertence e o maior destes dois valores. Para um conjunto de elementos, o coeficiente de Silhouette corresponde a média dos valores para a métrica dos elementos do conjunto. O domínio de valores resultantes é de -1 à 1, sendo o próximo de -1 indicativo de atribuições de agrupamentos erradas, próximo de 1 indicativo de atribuições de agrupamentos corretas e próximo de 0 indicativo de agrupamentos sobrepostos.

A equação 2.15 apresenta a fórmula para o cálculo do coeficiente de Silhouette para um elemento de um conjunto, onde:

- a representa a distância média entre o elemento em questão e os demais que pertencem ao mesmo agrupamento;
- b representa a distância média entre o elemento em questão e os demais que pertencem ao segundo agrupamento mais próximo dele.

$$s = \frac{b - a}{\max(a, b)} \quad (2.15)$$

O coeficiente de Calinski-Harabasz pode ser utilizado para avaliar um modelo de agrupamento de dados como a soma da dispersão entre os agrupamentos e a dispersão interna dos agrupamentos, onde a dispersão é calculada como a soma dos quadrados das distâncias envolvidas [9]. Valores mais altos se relacionam com agrupamentos que são densos e dissimilares dos demais, o que faz parte do conceito de um agrupamento.

A equação 2.16 apresenta a fórmula para cálculo do índice de Calinski-Harabasz, onde:

- $tr(B_k)$ representa a soma da dispersão entre os agrupamentos;
- $tr(W_k)$ representa a soma da dispersão interna dos agrupamentos; e
- n_E representa a quantidade de elementos do conjunto de dados.

$$s = \frac{tr(B_k)}{tr(W_k)} \cdot \frac{n_E - k}{k - 1} \quad (2.16)$$

A equação 2.17 apresenta a fórmula para cálculo da dispersão interna dos agrupamentos e a equação 2.18 apresenta a fórmula para cálculo da dispersão entre os agrupamentos.

$$W_k = \sum_{q=1}^k \sum_{x \in C_q} (x - c_q)(x - c_q)^t \quad (2.17)$$

$$B_k = \sum_{q=1}^k n_q (c_q - c_E)(c_q - c_E)^T \quad (2.18)$$

O coeficiente de Davies-Bouldin é capaz de representar a qualidade de um modelo de agrupamento de dados em função da separação entre os agrupamentos descritos [15]. O cálculo do coeficiente é feito pela média da similaridade entre agrupamentos, onde esta é uma métrica que relaciona a distância entre os agrupamentos com os tamanhos dos próprios agrupamentos. Desta forma, entende-se que, quanto mais próximo de zero, menor é a similaridade entre agrupamentos e, portanto, melhor é a qualidade do modelo.

A equação 2.19 apresenta a fórmula para cálculo de similaridade, onde:

- s_i representa o diâmetro do agrupamento i , ou seja, a média de distâncias entre os pontos pertencentes ao agrupamento i e o centróide do mesmo;
- s_j representa o diâmetro do agrupamento j ;
- d_{ij} representa a distância entre os centróides dos agrupamentos i e j .

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \quad (2.19)$$

A equação 2.20 apresenta a fórmula para cálculo do coeficiente de Davies-Bouldin, onde:

- R_{ij} representa a similaridade média entre os agrupamentos i e seu outro agrupamento mais próximo j .

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} (R_{ij}) \quad (2.20)$$

2.4 Análise de Variância e Teste de Tukey

Valores médios obtidos por diferentes conjuntos de diferentes condições são frequentemente comparados em estudos clínicos [40] e científicos [1]. Nas condições de distribuições normais e comparação entre duas médias, a avaliação da razão entre a diferença média e erros padrão da diferença média (teste t) é considerado adequada, mas quando a comparação acontece entre três ou mais médias, o método de comparação a ser considerado é a análise de variância (ANOVA) simples [40].

O resultado de uma ANOVA apresenta as somas das variâncias intergrupais, intragrupais e total; os graus de liberdade intergrupais, intragrupais e total; as médias dos quadrados das

variâncias intergrupais, intragrupais e total; e a estatística- F . A estatística- F é calculada pela razão entre a média do quadrado da variância intergrupal e a média do quadrado da variância intragrupal. A equação 2.21 apresenta a fórmula para o cálculo da estatística- F , onde:

- MQVE representa a média do quadrado da variância intergrupal; e
- MQVA representa a média do quadrado da variância intragrupal.

$$F = \frac{MQVE}{MQVA} \quad (2.21)$$

A média do quadrado da variância intergrupal é calculada pela razão entre a soma dos quadrados das variâncias intergrupais e a diferença entre a quantidade de grupos de dados analisados e 1. A fórmula 2.22 apresenta o cálculo para a média do quadrado da variância intergrupal, onde:

- SQVE representa a soma dos quadrados das variâncias intergrupais; e
- p representa a quantidade de grupos de dados analisados.

$$MQVE = \frac{SQVE}{p - 1} \quad (2.22)$$

A média do quadrado da variância intragrupal é calculada pela razão entre a soma dos quadrados das variâncias intragrupais e a diferença entre a quantidade total de observações analisadas e a quantidade de grupos de dados analisados. A equação 2.23 apresenta a fórmula para o cálculo da média do quadrado da variância intragrupal, onde:

- SQVA representa a soma dos quadrados das variâncias intragrupais;
- N representa a quantidade de observações analisados; e
- p representa a quantidade de grupos de dados analisados.

$$MQVA = \frac{SQVA}{N - p} \quad (2.23)$$

A partir do resultado de uma ANOVA, pode-se calcular o valor- p correspondente. Para tal, subtrai-se de 1 o resultado da aplicação da função de distribuição acumulada (FDA) da distribuição de Fisher-Snedecor (distribuição- F) parametrizada pela estatística- F e os graus de liberdade intergrupal e intragrupal, conforme encontrados por ANOVA. A equação 2.24 [70] apresenta a fórmula utilizada para verificação do valor- p , onde:

- FDA representa a função de distribuição acumulada da distribuição- F ;
- f representa o valor observado para a estatística- F ;

- $d1$ representa a quantidade de graus de liberdade intergrupais; e
- $d2$ representa a quantidade de graus de liberdade intragrupal.

$$p = 1 - FDA(f, d1, d2) \quad (2.24)$$

Quando uma ANOVA resulta em significância, isto indica que, ao menos, um dos tratamentos analisados se diferencia dos demais. Contudo, por si só, esta análise não informa o padrão de diferenças entre as médias analisadas, tampouco quais médias se destacam perante as demais [1]. Para descobrir tais especificidades, a ANOVA é comumente sucedida por técnicas de comparações par-a-par, como, por exemplo, o teste de diferença honestamente significativa (honestly significant difference, HSD) de Tukey [68, 1].

O teste de médias de Tukey calcula a diferença entre duas médias utilizando a distribuição de amplitude estudentizada (q) [1, 34]. A partir da distribuição- q é possível verificar a estatística- q exata da maior diferença entre um conjunto de médias oriundas da mesma população e, assim, todas as diferenças entre pares de médias são avaliadas em comparação com ela. A fórmula 2.25 apresenta o cálculo para estatística- q [34], onde:

- $\bar{X}_l$ representa a maior média encontrada entre os conjuntos de dados analisados;
- $\bar{X}_s$ representa a menor média encontrada entre os conjuntos de dados analisados; e
- $\sqrt{\frac{MQVA}{n}}$ representa a raiz quadrada da razão entre a média do quadrado da variância intragrupal e a quantidade de observações em cada conjunto de dados analisado.

$$q_r = \frac{\bar{X}_l - \bar{X}_s}{\sqrt{\frac{MQVA}{n}}} \quad (2.25)$$

Estabelece-se o valor crítico para estatística- q baseando-se na observação de uma tabela- q , que é construída a partir da distribuição- q . Esta observação é orientada pela quantidade de conjuntos de dados analisados, a quantidade de graus de liberdade da variância intragrupal e o limiar de significância a ser considerado, (comumente 0,05) [34].

Verifica-se, então, a estatística- q referente aos possíveis pares de médias dos conjuntos de dados analisados, conforme a equação 2.25. Caso a diferença entre a estatística- q referente a um par de médias supere o limiar definido pela estatística- $q_{crítica}$, considera-se que a diferença entre as médias é significativa.

2.5 Dissonância Cognitiva e Balanceamento Psicológico

Dentre outros aspectos dos processos mentais e comportamentais, a psicologia estuda fenômenos relacionados ao comportamento humano em consideração das relações interpessoais dos indivíduos e como elas influenciam suas percepções e ações sobre o ambiente e sobre outros indivíduos.

Este capítulo cita autores, respectivas teorias e conceitualizações, de interesse para o presente trabalho, que foram modelados computacionalmente no escopo do mesmo. Esta relevância se dá em que elas norteiam os processos de tomada de decisão dos agentes de software, em sua autonomia de movimentação pelo ambiente objetivando a aglomeração de agentes, conforme apresentado no capítulo 4.

Leon Festinger apresentou em 1957 a teoria da dissonância cognitiva [20], um fenômeno psicológico que se refere a situações onde há o conflito de atitudes, crenças ou comportamentos. Segundo a teoria, este tipo de situação causa desconforto mental, o que leva o indivíduo a alterar uma das atitudes, crenças ou comportamentos conflitantes, com vistas em reduzir tal desconforto e restaurar o equilíbrio psicológico.

A teoria sugere que as pessoas naturalmente têm o instinto de manter todas suas atitudes e comportamentos em harmonia, além de também evitar desarmonia, o que conceitua o princípio da consistência cognitiva. Na hipótese de inconsistência entre estes elementos, o que caracterizaria dissonância, os indivíduos buscariam alguma mudança de forma a diminuí-la. Contudo, é importante notar que a teoria afirma que o indivíduo busca mudanças com vistas em reduzir a dissonância cognitiva, apenas, mas não que as atitudes tomadas surtirão o efeito desejado na elevação de seu conforto psicológico.

A teoria do equilíbrio psicológico, conforme Heider (1957) [26], explica mudanças de atitudes dos indivíduos pela necessidade por equilíbrio psicológico. Segundo o autor, os indivíduos existem em constante busca pela manutenção da própria consistência cognitiva, resolvendo ou evitando conflitos entre seus valores e crenças e situações que encontram no ambiente onde estão inseridos.

Desta forma, apresenta um modelo para análise das relações perceptivas entre 3 entidades, onde elas se relacionam em caráter apreciativo, L , ou depreciativo, L' . No modelo, chamado de $P-O-X$, uma pessoa, P , e outra pessoa, O , são dotadas de capacidade cognitiva e outro elemento, X , pode ser um objeto de interesse ou outra pessoa igualmente dotada de cognições próprias. Há, ainda, a relação de unidade, U , onde denota-se que um objeto pertence ou é apenas associado diretamente a uma pessoa.

Havendo relações mútuas entre 3 entidades, encontra-se uma tríade de relações cuja análise revela o teor psicológico do ambiente ou região do ambiente em que as entidades se encontram, configurando a situação do mesmo. Segundo o autor, o teor psicológico de uma situação, equilibrado ou desequilibrado, pode ser calculado a partir do produto das relações, positivas ou negativas, entre as entidades que compõem a tríade.

Neste contexto, o autor afirma que tais situações podem ocorrer em 2 estados, sendo eles o equilibrado e o desequilibrado. No caso de um conjunto de relações apresentar teor psicológico de desequilíbrio, isto causaria tensão psicológica a uma entidade em questão e esta, instintivamente, atuaria de forma a diminuir tal desconforto através da modificação de uma ou mais relações mantidas por si em relação às demais envolvidas na tríade de relações ou, na impossibilidade disto, optaria pelo afastamento da situação.

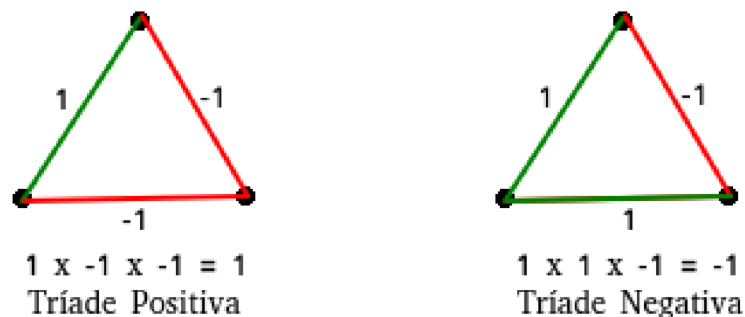
Consequentemente, segundo a teoria, conjuntos de relações cujo produto apresenta uma

situação de teor psicológico positivo não necessitariam de mudanças, enquanto conjuntos de relações que apresentam uma situação de teor psicológico negativo produzem desconforto cognitivo e pressionaria um interessado a realizar mudanças na sua percepção dos demais integrantes do conjunto de relações.

Cartwright e Harary (1956) [10] construíram uma extensão do modelo de Heider [26] na forma de uma representação da mesma fundamentada na teoria dos grafos e que permite a análise simultânea das relações entre vários indivíduos, em detrimento do modelo original, onde são consideradas 3 entidades. Isto foi realizado com vistas na aplicação em análise de redes sociais, em seu sentido amplo, onde haja um conjunto de indivíduos de qualquer natureza e um conjunto das respectivas relações interpessoais entre estes indivíduos.

Exemplos da simbologia adotada por Cartwright e Harary (1956) [10] são apresentados na figura 2.1. À esquerda, é apresentado um exemplo de uma tríade de relações que resulta em balanço positivo, visto que há uma relações positiva, destacada em verde, e duas relações negativas, destacadas em vermelho. À direita, é apresentado um exemplo de uma tríade de relações que resulta em balanço negativo, visto que há duas relações positivas, destacadas em verde, e uma relação negativa, destacada em vermelho.

Figura 2.1: Exemplos de tríades de relacionamento positivo e negativo, adaptado de Cartwright e Harary (1956) [10].



Uma contribuição da obra de Cartwright e Harary (1956) [10] para a teoria do equilíbrio psicológico foi a possibilidade de representar simultaneamente o conjunto de relações entre uma quantidade ilimitada de entidades. Contudo, esta generalização não foi aprofundada ao nível de situações específicas, por exemplo, em casos onde os conjuntos de indivíduos e respectivas relações compõem um grafo sinalizado completo, tal qual é proposta a modelagem das relações interpessoais por Heider [26]. Neste sentido, Flament (1963) [22] apresentou um teorema afirmando que um grafo sinalizado completo somente é balanceado se todas as tríades de relações que o compõe forem balanceadas.

Ao aprofundar o estudo de aplicações da teoria dos grafos em situações de relacionamentos interpessoais, Flament (1963) [22] afirmou que sua proposta anterior de atômica definir tais grafos como positivos ou negativos seria insuficiente, devido a que os diferentes indivíduos

têm percepções diferentes sobre tais situações. Desta forma, o autor propôs a classificação dos grafos de relações interpessoais de acordo com um grau de balanceamento.

Segundo o autor, o grau de balanceamento de um grafo pode ser calculado pela razão entre a quantidade de tríades positivas que compõem a situação e a quantidade total de tríades que compõem na situação. O resultado é um número fracionário de domínio entre 0 e 1 que permite a classificação do teor psicológico de um grafo de relações interpessoais com base em um dado quantitativo. Este conceito é sobremaneira apreciado no contexto do presente trabalho.

Embora uma das principais contribuições de Heider (1958) [26] para a teoria do equilíbrio psicológico seja baseada na definição de tríades de relações e, portanto, 3 entidades que compõem o conjunto de relações de uma tríade, o autor exemplifica situações onde haveriam apenas duas entidades envolvidas na análise do equilíbrio de relações, por exemplo, na reciprocidade de relações entre duas entidades, onde o teor psicológico seria sempre positivo ou sugerindo que uma possível ação em busca do equilíbrio psicológico seria se afastar de uma tríade em desconforto psicológico.

Isto abre precedente para o entendimento de que é possível inferir o teor psicológico a partir de uma situação composta por 2 entidades e, conseqüentemente, uma única relação entre as mesmas, e que situações nesta configuração sempre têm teor psicológico positivo ou balanceado.

Outra hipótese é encontrada onde uma situação apresentar todas relações em quantidade suficiente para formar tríades de mesmo teor psicológico, positivas ou negativas. Obedecendo o teorema de Flament [22], a situação teria o teor psicológico automaticamente definido como o único teor encontrado na situação, visto que tríades com todas relações de teor positivo resultam em situações positivas, enquanto a presença de relações unicamente negativas em uma situação impede que todas as tríades formadas assumir teor psicológico positivo.

Em continuação de seu trabalho, Festinger (1964) [21] revisitou sua teoria da dissonância cognitiva a fim de estudar como ela inerentemente afeta os processos de tomada de decisão na psicologia humana. Neste trabalho, o autor ponderou que, em face da necessidade por tomada de decisão, em que o indivíduo é capaz de identificar as vantagens e desvantagens de cada alternativa, reconhece que, ao tomar uma decisão, eliminará a apreciação dos benefícios das alternativas preteridas, da mesma forma que aceitará experienciar as desvantagens da opção tomada.

Assim, faz-se imprescindível diminuir a dissonância cognitiva causada pela necessidade de tomada de decisão, a fim de que o indivíduo tenha equilíbrio cognitivo ou, ainda, desconforto psicológico mínimo. Uma forma comum, segundo o autor, é aumentar o sentimento de atratividade pela opção escolhida e reduzir o sentimento de atratividade pelas opções preteridas, o que caracteriza o processo mental de ponderação de alternativas.

Antes disto, Brehm (1956) [8] estudou a relação entre a dissonância cognitiva e processos de tomada de decisão, realizando experimentos sociais que mostraram que indivíduos em situação de dissonância cognitiva tinham maior probabilidade de realizar a ponderação de alternativas na comparação com indivíduos em maior ou total conforto psicológico.

3 REVISÃO BIBLIOGRÁFICA

Esta revisão bibliográfica teve o objetivo de prover referencial de pesquisas relacionadas com o presente trabalho, de forma a evidenciar o ineditismo dos métodos introduzidos pelo mesmo.

O período abrangido pela pesquisa foi entre 2017 e 2019, sendo as consultas por publicações realizadas em uma plataforma² para busca por materiais relacionados a ciência da computação e áreas afins. Nela, são atualmente disponibilizados mais de 5 milhões de artigos relacionados a tecnologia, mais de 195 periódicos revisados por pares e mais de 1.800 conferências de abrangência global.

A busca foi realizada com a aplicação de critérios de inclusão e exclusão, em termos da presença de expressões específicas no título da obra ou no resumo do trabalho. Em alguns critérios também foram observadas a ordem em que as expressões foram apresentadas nos textos analisados. Em todos critérios, foram observadas flexões de gênero e número nas expressões utilizadas.

Nominalmente, foram incluídos trabalhos que contínham as expressões "algorithm for clustering" ou "clustering algorithm" ou "clustering method" no título do trabalho ou no seu resumo, sempre precedidos pelas expressões "novice" ou "new" ou "original" ou "novel", no mesmo elemento de metadados, com a máxima distância de 3 palavras. O critério de precedência imediata entre as expressões citadas foi preterido em vista da comum qualificação dos substantivos sucessores, conforme a utilização descrita, como, por exemplo, em "new data clustering algorithm" ou "novice k-means based clustering method".

Destes, foram excluídos aqueles textos cujos títulos ou resumos apresentavam as expressões "network" ou "text clustering" ou "image". Estes critérios de exclusão foram aplicados, respectivamente, a fim de evitar a seleção de trabalhos cujo tema fosse processamento distribuído ou paralelo em agrupamentos de computadores em redes, a aplicação específica do agrupamento de dados do agrupamento de texto e a aplicação específica de técnicas de agrupamento de dados no processamento de imagem.

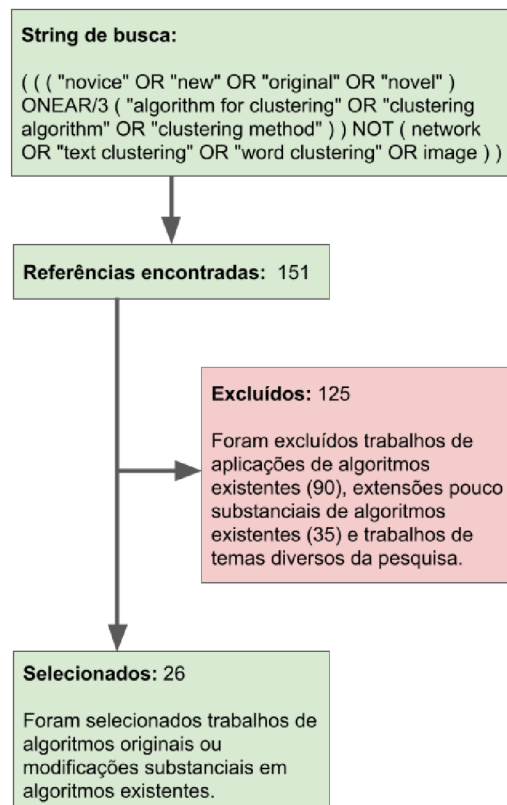
O total de trabalhos encontrados pela plataforma na busca descrita foi 151, dos quais 149 apresentados no idioma inglês, 1 em espanhol e 1 em turco. Destes, foram, ainda, descartados aqueles trabalhos que apresentavam apenas aplicações específicas de agrupamento de dados (90) ou melhorias em algoritmos já existentes e cuja publicação original extrapolara o escopo desta revisão (35). Ao final, foram selecionadas 26 obras para composição da revisão de trabalhos relacionados ao estudo presente. A figura 3.1 ilustra o protocolo de revisão realizado.

3.1 Trabalhos Relacionados

Esta seção apresenta trabalhos relacionados ao presente estudo. Tratam-se de algoritmos de agrupamento de dados pertencentes à revisão bibliográfica do mesmo, que compreende 26

²Plataforma IEEE Xplore. Disponível em <https://ieeexplore.ieee.org/>.

Figura 3.1: Protocolo de revisão realizado.



algoritmos publicados entre 2017 e 2019. O gráfico apresentado na figura 3.2 ilustra uma comparação visual em consideração da taxonomia apresentada na subseção 2.1.

A referida tipologia é baseada no mecanismo ou estrutura em que os diferentes algoritmos fundamentam o aprendizado, conforme [28], nominalmente: hierarquia, computação difusa, pseudo-centros, busca, grafos, malhas, densidade, modelos e sub-espacos. Destes, a presente revisão não apresenta algoritmos fundamentados na estrutura de malha ou seleção de sub-espacos, ambas voltadas a redução da complexidade computacional em dados volumosos.

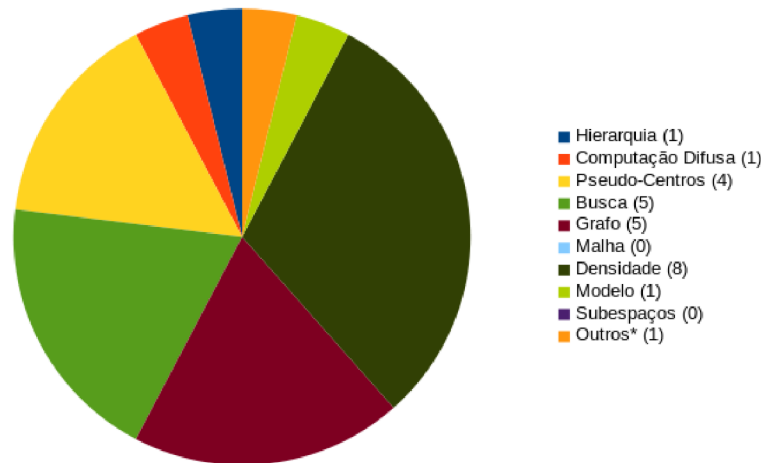
Há, ainda, um algoritmo apresentado nesta revisão que não foi encaixado na taxonomia devido a que, na verdade, compreende o aprendizado conjunto, onde abordagens diversas são aplicadas simultaneamente, com vistas em explorar as qualidades proeminentes de cada uma delas.

Optimal Leaf Optimization (OLO) Clustering

Rastin et al. (2017) [54] apresentaram uma proposta de agrupamento de dados relacionais, mas aplicável a dados vetoriais sem quaisquer perdas de qualidade nos resultados visto que o algoritmo possui a capacidade do cálculo das relações necessárias para continuação do processamento caso o conjunto de dados seja integralmente vetorial.

Caso as relações não sejam informadas ao algoritmo, o processamento inicia com a des-

Figura 3.2: Taxonomia de algoritmos de agrupamento de dados.



coberta das mesmas pelo cálculo de uma matriz de similaridade, onde consta um índice de similaridade para todas as possibilidades de pares dentre os pontos de dados contidos no conjunto processado. Na proposta, o critério de similaridade é a distância euclidiana entre dois pontos.

Esta matriz de similaridade é, então, reordenada pelo método de ordenação ótima de folhas [3], de forma a maximizar a soma das similaridades entre quaisquer pontos adjacentes na ordenação. A partir desta ordenação, é produzido um caminho hamiltoniano pelo grafo de pontos de dados, onde aqueles pertencentes ao mesmo agrupamento figuram próximos entre si e aqueles pertencentes a agrupamentos diferentes figuram distantes entre si na comparação com os pertencentes ao mesmo agrupamento.

A sequência de distâncias entre os pontos conforme o caminho hamiltoniano produzido na etapa anterior é dada como entrada em um algoritmo de processamento de sinais de frequência capaz de identificar saltos em sinais ruidosos. Estes saltos são, então, definidos como os locais do caminho hamiltoniano descoberto que definem os limites entre os agrupamentos, de forma a isolar entre si e os limites do próprio caminho hamiltoniano os pontos de dados pertencentes aos mesmos agrupamentos.

O algoritmo proposto teve o desempenho avaliado em testes sobre 4 conjuntos de dados artificiais e 6 conjuntos de dados reais, referentes a sequenciamento de proteínas. Na comparação de desempenho com os algoritmos DBSCAN, agrupamento espectral, propagação de afinidade e agrupamento hierárquico sobre os 10 conjuntos de dados utilizados, o algoritmo proposto obteve desempenho igual ou superior nos cálculos dos índices de Rand, informação mútua ajustada e métrica V.

Robust Dual Clustering with Adaptive Manifold Regularization (RDC)

Zhao et al. (2017) [81] propuseram um algoritmo de agrupamento de dados baseado no conceito de regularização de variedades de forma adaptativa (RDC) com vistas em otimizar a solução de problemas de agrupamento de dados em que ocorre alta dimensionalidade.

Ao incorporar este conceito na proposta, os autores o utilizaram especificamente no cálculo da matriz de afinidade que implica diretamente nos cálculos de pertencimento das instâncias aos agrupamentos. A regularização de variedades, no contexto do aprendizado de máquina, é uma técnica para aproveitamento do formato dos dados analisados, em termos de geometria euclidiana, para restringir as funções a serem aprendidas a partir do conjunto de dados.

Tratando-se de uma proposta de agrupamento espectral, foram realizados experimentos para verificação de desempenho do algoritmo RDC, onde fora comparado com os desempenhos dos algoritmos k-means, Normalized Cuts, Graph Regularized non-negative Matrix Factorization, Robust Manifold non-negative Matrix Factorization e Embedded Unsupervised Feature Selection.

Nos testes, foram utilizados 7 conjuntos de dados distintos cujas quantidades de instâncias variaram entre 24 e 9300, a quantidade de atributos entre 4 e 4434 e a quantidade de agrupamentos originais entre 3 e 20. As métricas de qualidade avaliadas foram as de acurácia dos agrupamentos, a informação mútua normalizada, a pureza dos agrupamentos e o tempo computacional de execução.

No tocante a acurácia dos agrupamentos, o algoritmo RDC teve melhor desempenho em todos cenários experimentados. O indicador de informação mútua normalizada apresentou desempenho superior para 39 dos 42 cenários experimentados. A pureza dos agrupamentos descobertos pelo RDC foi superior em 39 dos 42 cenários experimentados e o tempo computacional de execução do algoritmo inferior a média de execução dos demais algoritmos em apenas 5 dos 42 cenários experimentados.

Fitness Proportionate Share (FPS) Clustering

Yan et al. (2017) [78] apresentaram como problemas a necessidade que diversos algoritmos para agrupamento de dados possuem de requisitar parâmetros de entrada, juntamente com o conjunto de dados a ser processado, como, por exemplo, a quantidade de agrupamentos a ser descoberta e, também a incapacidade de tais algoritmos em detectar agrupamentos de densidades variadas.

Desta forma, os autores abordaram o agrupamento de dados baseando-se numa técnica de anichamento, comumente relacionado a algoritmos evolutivos, onde o processo da evolução por seleção de indivíduos por maior aptidão conforme Darwin é modelado computacionalmente.

Em algoritmos genéticos, é comum que o sistema convirja a uma solução ao passo que a diversidade populacional diminui, o que é conhecido como desvio genético. A manutenção da diversidade na população com base na distância entre os membros da mesma é chamada de anichamento. Isto se faz necessário para evitar que o algoritmo apresente uma solução ótima local como solução final em detrimento da solução ótima global.

Na proposta de Yan et al. (2017) [78], a técnica de anichamento aplicada isoladamente, sem aplicação do mecanismo evolutivo proposto pelos algoritmos genéticos na busca pelos centros dos agrupamentos, foi o FPS, um melhoramento do compartilhamento de aptidão tradicional

com vistas em evitar uma sensibilidade a soluções em nichos localmente ótimas.

O algoritmo identifica nichos com base nos picos de aptidão e aplica o FPS como estratégia de migração populacional entre os nichos identificados. A fusão de nichos ocorre após a identificação de todos eles e são realizadas até que suas conformações se estabilizem. Ao final, os picos de aptidão remanescentes são compreendidos como centros dos agrupamentos e cada ponto de dado atribuído a um agrupamento conforme uma métrica de distância, como a euclidiana.

Testes de desempenho foram realizados sobre 3 conjuntos de dados artificiais e comparação com o desempenho os algoritmos K-means, Subtractive e Fuzzy C-Means, onde o FPS-clustering apresentou superioridade na acurácia de classificação e índice de Kappa.

Uncertain Spectral Clustering (USC)

Wang et al. (2017) [74] afirmam que o agrupamento espectral é uma das formas mais eficazes para agrupamento de dados, onde é construída uma matriz de adjacência a partir de uma matriz de similaridade entre os pontos de dados. Em sua proposta, os autores estendem o método de agrupamento espectral para que haja suporte de dados incertos, baseando-se na semelhança entre distribuições de probabilidade.

Os autores explicam que a aplicação de métricas de semelhança entre distribuições de probabilidade ao problema de agrupamento de dados não era novidade, pois fora registrado emprego da divergência de Kullback-Leiber para a mesma finalidade. Ocorre que a aplicação da divergência de Kullback-Leiber tem eficácia comprovada sobre distribuições de probabilidade assimétricas, apenas, o que, segundo os autores, impossibilita a proposta de agrupamento espectral e, portanto, propõem aplicação da divergência de Jensen-Shannon (JS).

Além do conjunto de dados a ser analisado, o algoritmo necessita de parametrização referente a quantidade de agrupamentos a serem descobertos. Ele calcula uma matriz de divergência por JS, uma matriz de adjacência, uma matriz de admissão (Laplace) e um conjunto de autovetores referente a quantidade de agrupamentos a serem descobertos a partir da matriz de admissão calculada, o qual é utilizado por aplicação do algoritmo K-means como fundamento para alocação dos centros de agrupamentos.

Foram realizados 2 experimentos de agrupamento de dados sobre conjuntos de dados sintéticos. Para fins de análise de desempenho, os resultados foram comparados com os apresentados pelo algoritmo K-means no tocante as quesitos de correção média de classificação, média de informação mútua normalizada e média de tempo computacional de execução. Na comparação, o algoritmo proposto teve desempenho superior nos 3 critérios sobre os 2 conjuntos de dados analisados.

Data-Density-Aware Fuzzy Clustering (dFC)

Wang et al. (2017) [71] abordaram o problema do desbalanceamento, comumente encontrado em conjuntos de dados relacionados a biomedicina, e que posta um desafio aos métodos de agrupamentos de dados conhecidos.

O trabalho investigou os efeitos do desbalanceamento de dados em algoritmos de descoberta de agrupamentos por meio de lógica difusa e propõe um novo método de agrupamento de dados baseado em lógica difusa e sensibilidade a densidade de dados para solução ao problema.

O método inicialmente cria um particionamento temporário do conjunto de dados em diferentes agrupamentos com base nos picos de densidade dos dados apresentados. Então, o algoritmo iterativamente atualiza as informações dos centróides de cada agrupamento com base numa função própria de pertencimento, que é baseada na densidade de dados local para cada agrupamento, até que não seja necessária mais nenhuma atualização.

A avaliação de desempenho foi realizada numa comparação com o apresentado pelo algoritmo Fuzzy C-Means em testes sobre dois conjuntos de dados, respectivamente, de 5 e 2 agrupamentos, 2535 e 280 instâncias de dados e 2 e 6 dimensões. Para ambos casos, o algoritmo proposto, dFC, obteve melhor desempenho nos quesitos de acurácia nas classificações não supervisionadas e quantidade de iterações até encontrar a conformação dos agrupamentos. Em um dos testes, obteve comportamento inferior em termos de tempo computacional de execução até apresentação do resultado.

Sparse-Density Entropy (SDE)

Li et al. (2018) [41] propuseram um framework para o agrupamento de dados, capaz de realizar tarefas de pré-processamento, como o redimensionamento do conjunto de dados e a seleção de atributos com base na entropia de esparsidade, e a própria tarefa de agrupamento de dados, contexto sob o qual foi apresentado um algoritmo original, chamado entropia de densidade.

A abordagem apresentada para o algoritmo consiste em 11 etapas, iniciando na definição da quantidade de agrupamentos a serem descobertos a partir da soma de 1 e a raiz quadrada da quantidade de instâncias no conjunto de dados de entrada. O algoritmo procede à definição da distância total entre os pontos de dados, realiza os cálculos de densidade e vizinhos mais próximos para todos pontos de dados.

Também é calculada a fronteiridade de cada ponto de dado, de forma a identificar o conjunto de pontos fronteiriços que identificam os limites de cada agrupamento. Esta informação é refinada, a partir da identificação de pontos de dados ruidosos e é iniciada a definição dos agrupamentos a partir do ponto de menor densidade do conjunto, com agrupamento único. O agrupamento é estendido com a incorporação daqueles pontos que são alcançados pelo cálculo de densidade dos seus pontos agrupados. Com o fim da definição de cada agrupamento, um novo é iniciado a partir de algum ponto ainda não pertencente a quaisquer agrupamentos. Ao final, o algoritmo retorna os agrupamentos e o conjunto de pontos de dados ruidosos.

Na avaliação de desempenho do algoritmo, o framework SDE completo foi empregado em testes sobre 12 conjuntos de dados reais e sintéticos cujas quantidades de instâncias variaram entre 150 e 581012, de atributos entre 2 e 14409 e agrupamentos entre 2 e 25. Os demais algoritmos empregados na comparação de desempenho foram o DBSCAN, Statistical Informa-

tion Grid Clustering, K-means, Chameleon, Ng-Jordan-Weiss Clustering, Normalized Cuts e Self-Organizing Maps. No tocante a acurácia na classificação não supervisionada, o SDE teve o melhor desempenho sobre 10 dos 12 conjuntos de dados.

Zhong, Wang, Du e Tong (2018)

Zhong et al. (2018) [83] propuseram uma abordagem ao agrupamento de dados que utiliza de informações dos limites dos agrupamentos com base aprendizado por transferência. Segundo os autores, os pontos de dados que limitam os agrupamentos são definidos a partir da detecção das regiões do plano euclidiano de menor densidade de pontos de dados.

A avaliação de desempenho da proposta foi realizada pela comparação com os algoritmos K-means e propagação de afinidade sobre 7 conjuntos de dados artificiais bidimensionais e 5 conjuntos de dados reais de dimensionalidades que variaram entre 4 e 19 e volume de dados que variou entre 106 e 3970 instâncias. Os resultados mostraram desempenho superior do Boundary Information, embora os autores apresentem ressalvas quanto a conjuntos de dados caracterizados por desbalanceamento.

Local Linear Embedding Hybrid Kernel Affinity Propagation (LLK-HKAP)

Sun et al. (2018) [67] apresentaram um novo método de agrupamento de dados por propagação de afinidade baseado numa função de núcleo híbrido com integração linear local, abordando a problemática envolvida na extração de padrões de dados com altas dimensões e prover resultados em menor tempo computacional em comparação com algoritmos relacionados.

Inicialmente, o algoritmo para integração linear local é utilizado para reduzir a dimensão do conjunto de dados original. Após este redimensionamento, é aplicado o algoritmo de agrupamento de dados por propagação de afinidade original, que é baseado em uma métrica de similaridade de acordo com a função do núcleo híbrido utilizada.

Neste processo, um novo núcleo global, de grande capacidade de generalização, é definido enquanto uma função de núcleo híbrido que combina linearmente os núcleos global e Gaussiano é proposta para melhorar a capacidade de aprendizado do núcleo global. Logo, o novo núcleo híbrido é apresentado para definir uma métrica de similaridade e, a partir desta, possibilitar a criação de uma matriz de similaridade para o agrupamento de dados por propagação de afinidade.

Os autores provaram a eficiência do método de agrupamento de dados apresentado por testes realizados com este e outros algoritmos disponíveis na literatura, para comparação de desempenho, sendo todos variações e refinamentos do agrupamento de dados por propagação de afinidade com integração linear local e obtendo desempenho superior sobre os oito conjuntos de dados utilizados nos testes.

Clustering Based on Region Segmentation (CRS)

Chen, Chang e Zhao (2018) [12] propuseram um algoritmo de agrupamento de dados baseado em segmentação de regiões. Com o processamento dividido em 3 etapas, o algoritmo inicialmente segmenta o conjunto de dados em sub-regiões, procedendo, então ao cálculo de mesclabilidade para cada sub-região descoberta inicialmente, a fim de mesclar as sub-regiões em regiões maiores de acordo com alguma métrica, formando, então, os agrupamentos de dados.

A segmentação inicial do conjunto de dados em sub-regiões é realizada por cálculos de densidades locais e distâncias entre os pontos do conjunto de dados, de forma a definir os centros das sub-regiões. Nesta etapa, cada ponto é atribuído a um agrupamento correspondente a uma sub-região pela identificação do vizinho mais próximo, de uma determinada população próxima, com maior densidade e menor distância.

O cálculo de mesclabilidade entre as sub-regiões descobertas na etapa anterior é realizado a partir da detecção dos pontos de dados interseccionais entre as sub-regiões envolvidas no cálculo e, posteriormente, a detecção do cálculo do espaço interseccional entre as sub-regiões, de acordo com os pontos de dados também interseccionais. Caso haja um pico de densidade dentro do espaço interseccional que seja maior que o menor dos picos de densidade dentre as sub-regiões envolvidas no cálculo, então a mesclabilidade é verdadeira. Caso contrário, é dada como falsa.

Finalmente, o algoritmo mescla as sub-regiões de acordo com a mesclabilidade entre elas. O algoritmo CRS teve seu desempenho avaliado na comparação com o DBSCAN, DBSCAN com estrutura indexada e Statistical Testing Clustering sobre 13 conjuntos de dados cujas características variaram entre 178 e 10000 em termos de quantidade de instâncias, 2 e 40 na quantidade de agrupamentos a serem descobertos e 2 e 16384 em termos de dimensões analisadas. Dentre as 44 comparações de desempenho realizadas, o CRS obteve o melhor resultado em 40 delas.

Similar-to-Multiple-Point Clustering (MulSim)

Chen et al. (2018) [11] projetaram um algoritmo de agrupamento de dados baseado no distanciamento entre instâncias e robusto perante agrupamentos de diferentes formatos, densidades e população. De acordo com os autores, a abordagem se baseia na observação de que dois pontos de dados similares e que pertencem ao mesmo agrupamento tendem a possuir vizinhos similares e em comum, da mesma forma que dois pontos similares, mas pertencentes a agrupamentos diferentes, não compartilham vizinhos similares.

Considerando que o agrupamento de dados é baseado em alguma métrica de similaridade, a proposta apresenta uma nova distância que considera a posição ordinal em que dois pontos ocupam no conjunto ordenado de vizinhos mais próximos mutuamente. Segundo os autores, esta métrica, em detrimento de distâncias mais comumente utilizadas na literatura, como a euclidiana ou de manhattan, permite que o algoritmo supere o problema da descoberta de agrupamentos com diferentes densidades.

No momento da definição do pertencimento dos pontos de dados a um agrupamento, o algoritmo não somente considera a distância entre dois pontos, mas, sim, entre o ponto a ser agrupado e o conjunto de pontos que formam a vizinhança do ponto que pertence a um agrupamento. Somente na hipótese de o ponto a ser agrupado ser semelhante ao conjunto de pontos da vizinhança analisada, ele passa a fazer parte do mesmo agrupamento. Isto é realizado recorrentemente até que não hajam mais pontos de dados sem atribuição de agrupamento.

A fim de verificar o desempenho do algoritmo de agrupamento de dados pela descoberta de múltiplos pontos similares a um ponto (MulSim), foram realizados testes para comparação com outros algoritmos, especificamente DBSCAN, Ordering Points to Identify Cluster Structure Clustering, Fast Spatial Clustering via Binary Discretization, K-means, Clustering aLgorithm for Arbitrary ShaPed clusters e Clustering by Fast Search and Find of Density Peaks, tendo o MulSim superado os demais de acordo com as métricas de Rand ajustado e informação mútua normalizada em conjuntos de dados bidimensionais. Adicionalmente, foram realizados testes em conjuntos de dados de variadas dimensões, sendo o desempenho do algoritmo MulSim superado pelo apresentados pelos algoritmos K-means, Clustering by Fast Seach and Find of Density Peaks, Ordering Points to Identify Cluster Structure Clustering e DBSCAN em oportunidades distintas.

Closeness Measure based on Neighborhood Chain (CMNC) Clustering

Liang and Han (2018) [42] abordam a problemática da realização de agrupamento automático de dados em situações em que os agrupamentos corretos apresentam fronteiras complexas, como por exemplo, quando o formato dos agrupamentos corretos não é esférico ou a distribuição dos dados tem variações de densidade.

Para tal, os autores propuseram um algoritmo de agrupamento de dados baseado no encadeamento de vizinhança, uma nova métrica de proximidade diferente daquelas baseadas em distâncias geométricas, comumente utilizadas em algoritmos do tipo.

A métrica de proximidade baseada no encadeamento de vizinhança quantifica a proximidade entre dois pontos de dados baseando-se na dificuldade de alcance de um ponto a outro através de uma cadeia formada por seus vizinhos. Esta dificuldade de alcance é mensurada por duas métricas, o custo de alcance de vizinhança e a extensão do alcance da vizinhança.

O custo de alcance de vizinhança é definida com base no número que quantifica o mínimo de vizinhos necessários para estabelecer uma cadeia entre eles. A extensão do alcance de vizinhança, por sua vez, representa a maior distância entre dois pontos de dados adjacentes nas cadeias de vizinhos estabelecidas entre dois pontos.

Experimentos para verificar a eficácia da métrica de proximidade proposta foram realizados através de aplicação da mesma no algoritmo de agrupamento de dados K-means, utilizando-a como função de distância, ao invés da comumente utilizada função de distância euclideana e os resultados comparados com aqueles obtidos pelos algoritmos Ordering Points to Identify Cluster Structure Clustering, por propagação de afinidade e por detecção de picos de densidade. Sobre

os três conjuntos de dados distintos utilizados nos testes, a abordagem proposta teve resultado superior, com correção total na definição dos agrupamentos.

Ant Colony Optimization-Based Clustering Algorithm with Chaos (ACOCC)

Baseado no algoritmo tradicional de otimização por colônia de formigas (ACO) e sua aplicação numa abordagem de agrupamento de dados (ACOC), Yang et al. (2018) [79] propuseram um novo método de agrupamento de dados baseado em ACO, o agrupamento de dados baseado em otimização por colônia de formigas com caos (ACOCC).

A principal contribuição do trabalho foi a aplicação de uma função caótica nas fases de inicialização e atualização do feromônio das formigas da colônia. Com esta aplicação da função caótica na fase de inicialização do feromônio permite que as formigas sejam distribuídas nos diversos estados possíveis do sistema, enquanto a aplicação da função caótica na fase de atualização do feromônio pode causar distúrbios no algoritmo, de forma que as formigas explorem caminhos adicionais no ambiente, o que, conseqüentemente, evita que elas convirjam para soluções ótimas locais, ao invés de globais.

A eficácia do algoritmo proposto foi comprovada por experimentos de agrupamento de dados pelo ACOCC e pelo ACOC original sobre 4 conjuntos de dados de quantidades de agrupamentos, instâncias e dimensões diferentes, mostrando que o ACOCC tem maior facilidade em se afastar de soluções ótimas locais, em comparação com o ACOC.

Swarm Clustering Algorithm (SCA)

Zhu et al. (2018) [84] apresentaram uma nova proposta de agrupamento de dados por enxame de partículas. Os autores consideram que as aplicações então existentes da inteligência de enxames e algoritmos evolucionários no agrupamento de dados foram projetadas para que as partículas dos enxames representem os centros de agrupamentos.

No algoritmo de agrupamento de dados por enxame de partículas (SCA), cada ponto de dado do conjunto de dados de entrada é representado por uma partícula, de forma que a quantidade de partículas do enxame é igual a quantidade de instâncias do conjunto de dados. Todas estas partículas interagem dinamicamente com as partículas similares e se movimentam para regiões mais densas para formar agrupamentos.

Primeiramente, o enxame é iniciado com todas partículas em velocidade 0, melhores posições pessoais na posição inicial e têm suas densidades calculadas por estimativa de densidade do núcleo. O sistema procede à definição de uma partícula líder para cada partícula, podendo esta líder ser compartilhada por várias partículas. A definição do líder é feita pela escolha da partícula de maior densidade dentre as vizinhas de determinada partícula em cada iteração do processo.

A cada iteração, as partículas têm sua velocidade atualizada sob influência das melhores posições conhecidas por si, por um conjunto de partículas vizinhas e pela líder à qual está associada. Esta influência é definida pela métrica de peso de inércia. Isto é realizado até que

um critério de satisfação seja alcançado.

Finalmente, é calculado um raio de vizinhança para todas as partículas. Caso uma partícula se encontre dentro do raio de alcance de outra, elas pertencem ao mesmo agrupamento. Caso algum agrupamento apresente população inferior a determinada quantidade de partículas, o agrupamento é dissolvido e suas partículas atribuídas a outro agrupamento, mais próximo.

Foram realizados experimentos a fim de verificar a eficácia da abordagem, onde foram utilizados 5 conjuntos de dados, de diferentes populações, escalas e dimensões. Os algoritmos utilizados na comparação de desempenho foram DBSCAN, K-means, agrupamento hierárquico e aglomerativo e o Balanced Iterative Reducing and Clustering using Hierarchies, considerando as métricas de F, informação mútua normalizada e o índice de Rand ajustado. O SCA obteve melhor resultado em todos os testes na comparação com os algoritmos DBSCAN, agrupamento hierárquico e Balanced Iterative Reducing and Clustering using Hierarchies, sendo superado pelo algoritmo K-means sobre conjuntos de dados em que a quantidade de agrupamentos foi superior a 15.

Recursive Agent-Based Mixing (RAM)

Bhatti, Beck e Nedić (2019) [7] propuseram um algoritmo de agrupamento de dados baseado na teoria espectral de grafos, onde são estudadas propriedades de grafos em termos de autovetores e autovalores de matrizes associadas a um grafo, tais quais a matriz de adjacência ou a matriz oriunda do teorema de Laplace, por exemplo.

Os autores consideram a problemática da complexidade computacional do cálculo recorrente de autovalores, o que é comum em abordagens tradicionais do agrupamento de dados baseado na teoria espectral de grafos e que outros algoritmos que tentaram abordar o problema o fizeram em detrimento da qualidade dos agrupamentos descobertos.

Desta forma, procederam a utilização de um método que, ao invés de requerer explicitamente o cálculo recorrente de autovalores, calcula uma combinação linear de autovetores com pesos definidos por autovalores a partir de uma matriz de semelhança, que ocorre por servir como uma generalização do processo tradicional.

A estrutura resultante deste cálculo é utilizada na definição de movimentos dos pontos de dados para agregação de pontos semelhantes ou desagregação de pontos dessemelhantes em um plano euclidiano, processo o qual os autores chamaram de Recursive Agent-based Mixture. Finalmente, o algoritmo k-means é empregado a fim de descobrir os agrupamentos sobre a distribuição transformada.

O desempenho do algoritmo foi verificado em comparação com outros presentes na literatura utilizando a métrica de informação mútua normalizada. Os algoritmos utilizados na comparação foram o Shi-Malik, Fast-Approximate Spectral Clustering, Nystrom Clustering, Nystrom2 e Power Iteration Method Clustering. O algoritmo Recursive Agent-based Mixture obteve resultado superior aos demais algoritmos em 59 dos 63 cenários avaliados, onde os conjuntos de dados variaram em quantidade de agrupamentos no domínio de 4 a 20, tendo seu

desempenho superado pelo algoritmo Shi-Malik em 4 cenários, apenas.

Fast clustering by combining ENsembles of random projections with Scalable version of Improved Visual Assessment of Clustering Tendency (FensiVAT)

Rathore et al. (2019) [55] abordaram a problemática do agrupamento de dados massivos, considerando que muitos algoritmos então existentes são capazes de processar eficientemente conjuntos de dados de alta amostragem ou de alta dimensionalidade, mas comumente não são aplicáveis quando o conjunto de dados é massivo nos dois aspectos.

De forma a superar este problema, os autores apresentaram uma solução híbrida, referente a integração de um novo método de projeção aleatória conjunta, Improved Visual Assessment of Clustering Tendency e o método Maximin de amostragem aleatória. O algoritmo, chamado FensiVAT, realiza agrupamento rápido combinando as soluções de projeções aleatórias com uma versão escalável do algoritmo Improved Visual Assessment of Clustering Tendency.

A fim de verificar o desempenho do FensiVAT, foram realizados experimentos com este e os algoritmos K-means, Mini-Batch K-means, Clustering Large Applications, Clustering Using Representatives, clusiVAT, Gamma Region Density K-means, FastSpec. Projected Clustering e Random Projection-based Ensemble Clustering, sendo os 2 últimos projetados para analisar conjuntos de dados massivos. Os experimentos foram realizados sobre 2 conjuntos de dados sintéticos de 100000 instâncias e 1000 dimensões e 6 conjuntos de dados reais, de forma a medir o tempo computacional de execução e acurácia dos agrupamentos. O algoritmo FensiVAT se mostrou superior na comparação de desempenho com os demais algoritmos testados em ambos aspectos.

Belief Peaks Evidential Clustering (BPEC)

Su e Denoeux (2019) [66] apresentaram um algoritmo de agrupamento de dados baseado na teoria das evidências, conforme Dempster e Schafer (1976). A referida teoria trata de um mecanismo genérico para tratamento de incertezas, envolvendo conceitos de probabilidade, possibilidade e probabilidade imprecisa.

A proposta utiliza o conceito de picos de crenças para identificação dos centros dos agrupamentos. Esta noção se assemelha ao conceito de picos de densidade, oriundo dos métodos de agrupamento de dados baseado em picos de densidade onde é definida uma função de massa na descoberta dos pontos de dados que posteriormente assumem o papel de centro dos agrupamentos.

No caso da proposta de Su e Denoeux (2019), a descoberta dos pontos de dados mais proeminentes ocorre pela definição de um grau de crença aos pontos de dados a partir da definição de uma função de crença. O fundamento da abordagem está na premissa de que todos os pontos de dados na vizinhança de cada ponto de dado apresentam evidências de que tal ponto de dado seja um centro de agrupamento.

Neste contexto, quaisquer pontos de dados que apresentarem maiores crenças do que seus

vizinhos e estiverem localizados em certa distância de outros centros de agrupamentos são definidos como centros de agrupamentos. Finalmente, os pontos de dados são associados a um centro de agrupamento em acordo com uma métrica de distância adaptativa, de forma a permitir a descoberta de agrupamentos em formatos diferentes. O algoritmo apresentado teve a nomenclatura Belief-Peaks Evidential Clustering.

O desempenho da abordagem foi avaliado em comparação com algoritmos de agrupamento de dados baseado em picos de densidade, nominalmente o Adaptive Density Peaks Clustering based on K-Nearest Neighbors, Fuzzy K-Nearest Neighbors Density Peaks Clustering, Density Peaks Clustering based on K-Nearest Neighbors e Density Peaks Clustering. Na comparação pelo índice de Rand ajustado, o BPEC superou ou igualou os demais algoritmos em 54 oportunidades, sendo superado em apenas 2 pelo algoritmo Fuzzy K-Nearest Neighbors Density Peaks Clustering. Os conjuntos de dados utilizados nos testes variaram na quantidade de instâncias entre 150 e 7500, atributos entre 2 e 1024 e agrupamentos entre 2 e 50.

Os autores apresentaram, ainda, uma versão do algoritmo Density Peaks Clustering meramente adaptada para a detecção de picos de crença ao invés de picos de densidade. Esta adaptação apresentou desempenho igual ou superior na comparação estrita com o algoritmo Density Peaks Clustering em 12 de 14 testes, sendo superada em apenas 2, segundo o índice de Rand ajustado, embora que inferior a abordagem principal, BPEC.

Hybrid and Parameter-free Clustering Method (HPFCM)

Shao et al. (2019) [62] afirmam que muitos algoritmos para agrupamento de dados são projetados para desempenho em conjuntos de dados relativamente pequenos, o que os torna ineficientes quando a quantidade de dados a serem analisados é massiva. Adicionalmente, os autores apontam a necessidade de muitos algoritmos por parâmetros adicionais para a realização da tarefa de agrupamento como prejudicial ao desempenho do mesmo devido a que tais parâmetros podem ser difíceis de obter em aplicações reais.

Desta forma, os autores propuseram um algoritmos para agrupamento de dados chamado Hyperparameter-free Clustering Method, que é capaz de agrupar dados massivos rapidamente sem conhecimento prévio da quantidade de agrupamentos a ser descoberta.

O método é baseado numa variante da amostragem aleatória Maximin em que são selecionados dados que estejam na máxima distância entre si no domínio do conjunto de dados analisado, de forma que a amostra dos dados retenha a geometria do conjunto de dados original. Na extensão, os autores adicionaram uma etapa de seleção aleatória de dados anteriormente ao algoritmo de amostragem aleatória Maximin, com vistas em reduzir a complexidade computacional exigida pelo método original.

Esta amostragem de dados é utilizada pelo algoritmo Efficient Visual Assessment of Clustering Tendency, para descoberta das tendências de agrupamentos a partir de uma imagem de profundidade e uma árvore de expansão mínima, construídos a partir da interpretação do conjunto de dados de entrada. A extração de blocos de pixels é uma proposta para automação

da descoberta de blocos de pixels pretos na imagem processada pelo algoritmo, de forma que cada bloco descoberto na imagem representa um agrupamento presente na amostra de dados analisada.

Conhecendo a quantidade de agrupamentos a serem descobertos a terceira etapa do processamento consiste na aplicação de uma nova árvore de expansão mínima para descoberta dos agrupamentos presentes na amostra de dados analisada e, finalmente, os resultados são estendidos ao restante do conjunto de dados original.

Os autores avaliaram o comportamento de sua proposta na comparação com os algoritmos HPFCM, Scalable Improved Visual Assessment of Clustering Tendency, Fuzzy C-Means e K-means e obteve desempenho superior em testes sobre 9 conjuntos de dados massivos distintos que variaram na quantidade de instâncias entre 100.000 e 500.000 e a quantidade de agrupamentos originais entre 3 e 5. O algoritmo também teve desempenho superior em 9 testes sobre dados altamente dimensionados, onde a quantidade de atributos variou entre 100 e 500, sobre 7 conjuntos de dados de pixels de imagens de números escritos a mão, cujas dimensões variaram entre 6 e 784.

Clustering Feature Tree-Based Effective Multiple Range Queries-Based Clustering (CF+-ERC)

Ryu, Jung e Pramanik (2019) [58] propuseram uma nova abordagem ao problema de agrupamento de dados com base na redução de complexidade a partir do resumo do conjunto de dados original. Os autores pontuam que outros métodos de descoberta de agrupamentos por hierarquias, como o algoritmo Balanced Iterative Reducing and Clustering using Hierarchies, têm a capacidade de construir uma árvore de características, mas utilizam apenas as folhas da árvore para definição dos agrupamentos.

Outrossim, o trabalho propôs o algoritmo de efetivo agrupamento baseado em busca de múltiplo alcance (CF+-ERC), que é capaz de obter ganho de informação relevante para tarefas de agrupamento de dados a partir da total estrutura de uma árvore de características criada quando do agrupamento por hierarquias. Os autores também apresentaram uma estrutura otimizada para a árvore de características, que aprimora a seleção do atributo para divisão das partições.

Na conclusão, os autores consideraram sua proposta para agrupamento de dados com desempenho competitivo na comparação com um algoritmo de agrupamento hierárquico e aglomerativo, tendo encontrado resultados superiores, semelhantes e inferiores nos quesitos pureza dos agrupamentos, revocação e tempo computacional de execução dependendo das condições dos padrões dos agrupamentos, quantidade de agrupamentos descobertos, quantidade de instâncias e de dimensões do conjunto de dados e origem sintética ou real do conjunto de dados.

Density Peaks Clustering using Symmetric Neighborhood Relationship (DPC-SNR)

Wu et al. (2019) [77] tratam de uma problemática inerente ao algoritmo Density Peaks Clustering original, que, segundo os autores, possibilita a seleção inicial dos centros dos agrupamentos sem ações corretivas no decorrer da execução do algoritmo, causando que o mesmo não seja capaz de identificar corretamente agrupamentos com diferentes densidades e, em consequência, prejudique a qualidade dos agrupamentos descobertos.

Outrossim, os autores propõem um método robusto de agrupamento de dados por picos de densidade em que é construído um grafo de vizinhança simétrico para todos pontos de dados (DPC-SNR), baseando-se no algoritmo de descoberta de vizinhos mais próximos tradicional e aplicação reversa da detecção de vizinhos mais próximos para distinção de picos de densidade locais para cada ponto de dado.

Após a construção do grafo de vizinhança simétrico, os centros iniciais dos agrupamentos são estimados sobre os picos de densidade e os agrupamentos semelhantes são agregados com base no grafo, resultando em que todos pontos de dados sejam alocados em um dos agrupamentos descobertos.

Experimentos para verificação de desempenho foram conduzidos, bem como a comparação com o executado pelos algoritmos DBSCAN, Density Peaks Clustering, Adaptive Density Peaks Clustering based on K-Nearest Neighbors e Density Peaks Clustering based on K-Nearest Neighbors and Principal Component Analysis sobre 18 conjuntos de dados sintéticos e reais, cujas quantidades de instâncias variaram entre 150 e 4177, atributos entre 3 e 36 e agrupamentos originais entre 2 e 6. Os resultados obtidos mostraram que o algoritmo proposto pode ter desempenho superior aos melhores desempenhos dos demais algoritmos aplicados nos experimentos em casos específicos altamente dependentes da parametrização.

Mortezanezhad e Daneshifar (2019)

Mortezanezhad e Daneshifar (2019) [48] propuseram um algoritmo de agrupamento de dados baseado na proposta evolutiva do algoritmo genético, que compreende uma heurística de busca inspirada na teoria da evolução natural Darwiniana e a lei da sobrevivência dos indivíduos mais aptos.

A proposta explora a aleatoriedade natural do algoritmo genético, de forma que cada cromossomo gerado representa o conjunto de centros aleatórios para uma quantidade de agrupamentos também aleatória, embora dentro de um intervalo.

Após a geração da população inicial, em cada geração são realizadas etapas características do algoritmo genético. Elas compreendem o cálculo da função de aptidão, aplicação dos operadores de cruzamento, aplicação dos operadores de mutação e uma busca local, esta baseada no comportamento do algoritmo K-means para atribuição dos agrupamentos para cada instância de dados sendo agrupada.

Isto culmina em que os cromossomos de melhor aptidão são enviados para a próxima geração, onde o processo se repete até que um limite de gerações seja alcançado. Ao final, o cromossomo

de melhor aptidão entre todas gerações é a solução apresentada pelo algoritmo.

Para verificação de desempenho, o algoritmo foi executado sobre 4 conjuntos de dados distintos, característicos pelo formato esférico dos agrupamentos e desbalanceamento entre as classes. Em comparação com o algoritmo Efficient Grouping Genetic Algorithm for Data Clustering, a proposta apresentou desempenho superior nos 4 casos estudados. Os autores propuseram, ainda, a aplicação de mecanismos de seleção de atributos e redução de dimensionalidade em trabalhos futuros.

Fully-informed Particle Swarm (FIPS)-based Clustering

Mansour e Ahmadi (2019) [46] introduziram um novo algoritmo para agrupamento de dados baseado em otimização por enxame de partículas totalmente informado (FIPS). Segundo os autores, o algoritmo de otimização por enxames de partículas totalmente informado, fundamentado no algoritmo de otimização por enxame de partículas (PSO) original, propôs a aplicação de diferentes topologias de vizinhanças.

Neste estudo, é proposto que três topologias de vizinhanças sejam empregadas para a finalidade de agrupamento de dados, sendo elas as topologias total, em anel e por distância. Destas, apenas a última não havia sido utilizada em propostas de solução de problemas de otimização, cuja experimentação materializa a principal contribuição do trabalho, o algoritmo FIPS-D.

A topologia total representa o comportamento original do algoritmo de PSO, onde todas partículas se comunicam entre si para determinar a cognição social que parcialmente influencia os movimentos das partículas. A topologia em anel considera a ordem em que os dados se encontram e a informação de quantas instâncias imediatamente anteriores e posteriores, de acordo com esta ordenação inicial, devem ser consideradas a vizinhança (outras partículas com quem a comunicação é permitida) de cada partícula; este comportamento proporciona que as informações se propaguem mais lentamente no grafo que representa as partículas.

A topologia de distância é baseada na distância entre as partículas do enxame e a vizinhança de cada partícula consiste em determinada quantidade de outras partículas mais próximas; diferentemente de outras topologias, esta propõe dinamismo nas vizinhanças de cada partícula, de forma que as distâncias entre elas mudam a cada iteração do algoritmo e, conseqüentemente, as vizinhanças também mudam.

O desempenho do algoritmo foi verificado em comparação com outros quatro algoritmos de agrupamento de dados que utilizam a otimização por enxame de partículas como mecanismo fundamental. Sobre oito conjuntos de dados distintos, o FIPS-D teve desempenho superior sobre sete deles, sendo superado pela abordagem baseada na topologia de vizinhanças em anel em uma ocasião.

First-Integer-Nighbor-based Clustering Hierarchy (FINCH)

Sarfraz, Sharma e Stiefelhausen (2019) [59] propuseram um algoritmo de agrupamento de dados baseado na premissa de que a única informação necessária para descobrir agrupamentos de

dados de forma eficaz é o vizinho mais próximo de cada ponto de dado e esta é a única estatística necessária para a descoberta de cadeias e ligações dentre o conjunto de dados analisado.

Na proposta, em detrimento da manutenção de uma matriz de distâncias entre os pontos, é criada uma matriz de adjacência a partir de uma equação que verifica duas condições entre dois pontos quaisquer: ser o vizinho mais próximo ou compartilhar o vizinho mais próximo. Somente em caso positivo, os dois pontos são adjacentes. Finalmente, o algoritmo possibilita a mesclagem das partições descobertas para definição dos agrupamentos de dados.

O desempenho do algoritmo foi verificado na comparação com os métodos dos algoritmos K-means, Balanced Iterative Reducing and Clustering using Hierarchies, Spectral Clustering, agrupamento hierárquico e aglomerativo, Sparse Subspace Clustering, Multi-view Low-rank Sparse Sub-space Clustering e Robust Continuous Clustering. Em termos do índice de informação mútua normalizada, o algoritmo proposto, FINCH, obteve resultados superiores aos demais algoritmos verificados em testes sobre 4 conjuntos de dados distintos, sendo superado na utilização sobre apenas 1 conjunto de dados pelo algoritmo Robust Continuous Clustering.

Novel Insect Pollination Clustering (NIPC)

Qu, Wu e Li (2019) [52] propuseram um algoritmo de agrupamento de dados baseado em inteligência de enxames e inspirado no processo biológico da polinização de plantas mediada por insetos, aderindo à premissa de que plantas da mesma espécie tendem a se agregar em termos de localização espacial. Segundo os autores, a natureza age em suporte a um processo evolutivo em que plantas próximas apresentam grãos de pólen coespecíficos.

O algoritmo, inicialmente, normaliza o conjunto de dados de forma que cada ponto de dado representa uma planta de ciclo anual e cada atributo uma característica das plantas que podem ser utilizadas para fins de agrupamento. O ambiente é, então, virtualmente populado com agentes cujo comportamento modela o de abelhas polinizadoras nas ações de percepção do ambiente, memorização, forrageamento, extração de pólen e deposição de pólen.

Cada abelha escolhe uma planta aleatoriamente para ser seu alvo de polinização inicial. Caso a probabilidade de sobrevivência da planta para o próximo ciclo seja superior a um limiar, sua posição não muda. Caso contrário, ela é realocada de acordo com a semelhança entre seus atributos e os de outras plantas presentes na memória do agente polinizador. O agente move-se para outra planta aleatoriamente, e o processo se repete até que seja atingida determinada quantidade de iterações.

Para fins de verificação de eficácia do método, o algoritmo NIPC teve o desempenho comparado com o Adaptive Artificial Ants Clustering e FlyAntClass em tarefas sobre conjuntos de dados de testes cujas características variaram em termos de quantidade de instâncias de dados, de 150 a 600, agrupamentos, de 3 a 5 e dimensões, de 2 a 13. As comparações foram realizadas em termos de compactação e separação entre os agrupamentos descobertos além da métrica F, onde o algoritmo NIPC teve o desempenho superado em apenas uma ocasião, pelo algoritmo Adaptive Artificial Ants Clustering na observação da métrica F em um dos testes.

Internally Wighted Graph Cut (IWGC) Clustering

Zhao e Lian (2019) [82] abordam um problema inerente a aplicações do aprendizado conjunto em tarefas de agrupamento de dados, que é a suposição de que todos os componentes do aprendizado conjunto têm igual influência na tomada de decisão pelo resultado final dos agrupamentos. Segundo os autores, isto tem efeito negativo na qualidade do modelo final nos casos em que componentes de baixa qualidade têm influência sobre o mesmo.

Ademais, outra situação é que mesmo considerando as qualidades dos modelos individualmente ao atribuir a influência do mesmo na qualidade do modelo final, é possível que parte dele, como por exemplo um ou mais agrupamentos, sejam retratados com alta qualidade dentro deles, apesar da baixa qualidade do modelo componente como todo.

Desta forma, objetivo do trabalho foi um método de agrupamento de dados que, ao utilizar de aprendizado conjunto, leva em consideração a diversidade de agrupamentos na estrutura interna de cada modelo componente do conjunto, melhorando o desempenho do modelo final.

Inicialmente, a proposta adota 5 algoritmos de agrupamento de dados na geração de componentes distintos, sendo eles K-means, Gaussian Mixture Models Clustering, Means-Shift Clustering, agrupamento hierárquico e aglomerativo e DBSCAN. Após a aquisição de diversos modelos componentes, aqueles de baixa qualidade são descartados, sendo os remanescentes são avaliados no tocante a confiabilidade dos agrupamentos apresentados. Finalmente, uma função de consenso é aplicada a fim de alcançar o modelo final.

Experimentos foram realizados para avaliar o desempenho do método proposto em comparação com outros métodos de agrupamento de dados por aprendizado conjunto. Seis conjuntos de dados, com dimensões de agrupamentos variantes de 2 a 14, quantidade de instâncias de 400 a 45211 e atributos de 17 a 103. Os resultados mostraram superioridade de desempenho do método proposto quando em comparação com os algoritmos de agrupamento de dados utilizados na criação dos modelos componentes do aprendizado conjunto. Também houve superioridade de desempenho na comparação com outros métodos de aprendizado conjunto em exceção de 2 das 30 ocorrências, onde algum método de aprendizado conjunto apresentou desempenho igual ou superior.

Transfer Clustering based on Gaussian Mixture Model (TGMM)

Wang et al. (2019) [72] abordaram a ineficácia do agrupamento de dados pelo método Gaussian Mixture Models quando os dados são escassos. Outrossim, propuseram um método baseado em aprendizado por transferência para aproveitar informações de uma tarefa base, anterior, para impactar na qualidade dos agrupamentos descobertos numa tarefa alvo, futura.

No método, primeiramente o modelo de mistura Gaussiana tradicional é aplicado na tarefa base sobre determinado conjunto de dados, de forma a extrair as médias e covariâncias de cada distribuição gaussiana, conforme proposta do modelo. Os dados da tarefa alvo são, então, agrupados sob influência dos parâmetros extraídos pela tarefa base.

Em comparação com o método de agrupamento pelo modelo de mistura de gaussianas

sem aplicação do aprendizado por transferência, a proposta mostrou desempenho superior no indicador de Rand em experimentos com 3 conjuntos de dados sintéticos distintos. Os autores apontaram, ainda, a necessidade de a quantidade de clusters oriundos das tarefas base e alvo serem as mesmas como uma limitação da proposta.

Robust Density Clustering Algorithm Based on Gravity Peak (RDC-GP)

Zhang et al. (2019) [80] reconhecem a evidência da abordagem ao problema de agrupamento de dados baseada em detecção de picos de densidade, mas criticam o fato de o método ser originalmente ineficaz ao distinguir os centros de agrupamentos em conjuntos de dados de conformação uniforme e heterogênea.

Além disto, algoritmos de agrupamento de dados baseado em picos de densidade apresentam dificuldades ou incapacidade de distinguir agrupamentos cujas áreas se sobrepõem. Com vistas em tratar estas problemáticas, os autores propuseram o algoritmo de agrupamento de dados robusto baseado em picos de gravidade (RDC-GP).

O primeiro problema, segundo os autores, pôde ser resolvido pela introdução do conceito de gravidade, fazendo uma analogia entre a quantidade de instâncias de dados em cada agrupamento e a massa de um objeto, que equaciona a força gravitacional na área da física. O segundo problema pôde ser resolvido por um método de detecção de inclinação gravitacional, capaz de identificar os dados limítrofes de cada agrupamento.

A eficácia do método foi avaliada em experimentos em 10 conjuntos de dados distintos, cuja quantidade de agrupamentos variaram entre 2 e 15, quantidade de instâncias entre 72 e 2048 e dimensões entre 2 e 34. Os desempenhos obtidos nestes testes foram comparados com os resultados encontrados em experimentos semelhantes com outros 7 algoritmos de agrupamento de dados encontrados na literatura.

A métrica avaliada foi a acurácia de classificação conforme os agrupamentos originais dos conjuntos de dados. Para 5 dos 8 conjuntos de dados testados, o RDC-GP apresentou o melhor resultado, e em outras 2 ocasiões apresentou o segundo melhor desempenho.

3.2 Síntese da Revisão

Este capítulo apresentou os resumos dos 26 trabalhos de publicação de novos algoritmos ou abordagens ao problema do agrupamento de dados do período de 2017 à 2019. Todos os citados são relacionados com a presente proposta na investigação da tarefa de mineração de dados do agrupamento de dados (clustering) e evidenciam a relevância do assunto e seus tópicos na atualidade. Porém, no contexto de relação com a metodologia de agrupamento de dados proposta na presente obra, destacam-se os algoritmos ACOCC [79], SCA [84], FIPS [46] e NIPC [52], devido à aplicação de mecanismos de otimização de enxame de partículas.

Algoritmos de agrupamento de dados que utilizam deste mecanismo ou estendem o Particle Swarm Clustering (PSC) original [13], naturalmente, herdam sua fundamentação principal que,

segundo os autores, é o conceito da dissonância cognitiva. No caso do algoritmo PSC e dos demais citados na revisão, as cognições consideradas pelos indivíduos do sistema são a própria e a social. Quando elas divergem na indicação da direção do movimento o indivíduo, surge a dissonância cognitiva e a busca por equilíbrio.

Na presente proposta, tal dissonância é criada a partir do momento em que um indivíduo observa uma facilidade de agrupamento nas diversas direções as quais enxerga, mas também o teor de equilíbrio nas relações entre os indivíduos nas regiões. Caso uma direção de movimento que propicia maior facilidade de agrupamento tenha, também, uma situação de teor psicológico negativo, surge a dissonância cognitiva e a busca por outro tipo de movimentação, visando o equilíbrio. Assim, fica exposta a relação dos algoritmos destacam-se os algoritmos ACOCC [79], SCA [84], FIPS [46] e NIPC [52] com a presente proposta: a fundamentação das tomadas de decisão dos indivíduos do sistema no conceito de dissonância cognitiva.

4 METODOLOGIA

Este capítulo apresenta o algoritmo de agrupamento de dados desenvolvido, detalhando seu comportamento e a influência dos conceitos relacionados à dissonância cognitiva nas tomadas de decisão. O algoritmo trabalha, dentre outros conceitos, com o de sistemas multiagentes e o processo de agrupamento de dados é realizado de acordo com a movimentação dos agentes dentro do sistema, cujas ações no contexto são em busca de aglomeração. O conjunto de dados utilizado nas ilustrações do comportamento do algoritmo foi composto, para fins da exemplificação, por 50 instâncias de pontos bidimensionais. As ilustrações retratam os pontos de dados dispostos num plano euclidiano bidimensional. As equações apresentadas modelam matematicamente as heurísticas envolvidas nas tomadas de decisão.

4.1 Pré-processamento e Inicialização dos Agentes

Inicialmente, são instanciados agentes de software autônomos para representação individual de cada instância contida no conjunto de dados de entrada. Neste momento, caso o conjunto de dados de entrada apresente quantidade de dimensões superior a 2, é realizada a redução de dimensionalidade do mesmo. Isto possibilita que os agentes existam e se movimentem em um ambiente bidimensional, em detrimento de um ambiente de dimensionalidade múltipla.

A técnica de redução de dimensionalidade utilizada neste estudo foi a análise de componentes principais (PCA). Ela foi elencada por ser a mais adotada em estudos desta natureza [17], o que facilita o comparativo com outras abordagens. Contudo, o algoritmo proposto está preparado para utilizar outras técnicas de redução de dimensionalidade.

Também é possibilitada a parametrização de aspectos das tomadas de decisão que os agentes realizam no contexto do sistema. Nominalmente, são eles: o tamanho do passo que os agentes são capazes de se movimentar em cada iteração, o tamanho do raio do círculo de percepção em torno dos agentes que indica o espaço no qual eles são capazes de detectar a presença de outros agentes, a quantidade de setores em que o círculo de percepção dos agentes é divididos, o limiar de satisfação cognitiva atribuído aos agentes do sistema, que pauta a definição do teor psicológico de uma situação interagente como positivo ou negativo, e a quantidade de agentes vizinhos mais próximos com os quais cada agente estabelece relações positivas.

Cada agente tem inicializadas as respectivas propriedades de posição multidimensional original no plano euclidiano, posição multidimensional atual no plano euclidiano, uma string de identificação individual, uma string de identificação do agrupamento de agentes ao qual pertence, o raio do círculo de percepção em torno de si em que é capaz de detectar a presença de outros agentes, uma cor aleatória que pode ser utilizada para representação visual e um conjunto inicialmente vazio para o armazenamento temporário dos demais agentes presentes em seu campo de visão, que é repopulado em cada iteração do algoritmo para fins de otimização computacional.

4.2 Posicionamento e Movimentação dos Agentes

A figura 4.1 apresenta a disposição de 50 agentes em um plano euclidiano bidimensional, onde cada agente representa um ponto de dado. Para fins de exemplificação do comportamento dos agentes no sistema, é destacado o ponto aproximadamente localizado nas coordenadas 1.2 e 3.0, em vermelho. A figura 4.2 ilustra como o agente identifica seu círculo de percepção da presença de outros agentes, no caso do exemplo um raio de 1.25 unidades de distância, e a figura 4.3 ilustra como, posteriormente, secciona seu círculo de percepção, no caso do exemplo em 8 segmentos.

Figura 4.1: Distribuição de agentes no ambiente 2D.

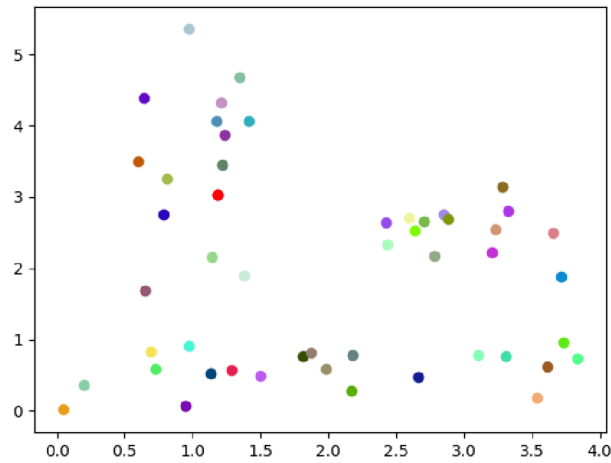
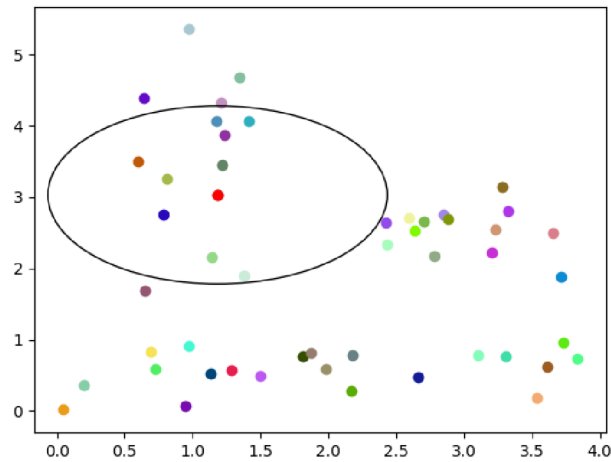
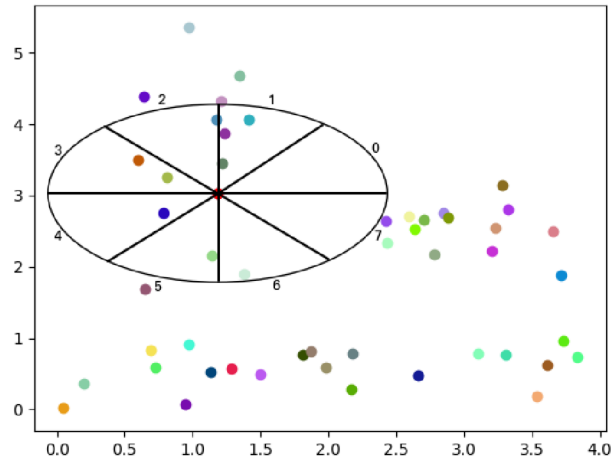


Figura 4.2: Círculo de percepção de agente em destaque.



Ao início de cada iteração, a ordem em que os agentes definem suas movimentações no ambiente é aleatorizada a fim de evitar que os fenômenos observados sejam relacionados a ordenação das instâncias no conjunto de dados de entrada.

Figura 4.3: Círculo de percepção do agente em destaque dividido em seções.



O sistema constrói um grafo de relações interagentes para todos aqueles presentes no sistema. Tais relações foram projetadas de forma que cada agente mantém uma relação positiva com todos os outros agentes cuja distância é igual ou inferior à distância entre si e o n -ésimo agente em ordem de distância deste mesmo. Com todos os outros agentes, o agente em questão estabelece uma relação negativa. Esta quantidade de agentes mais próximos com os quais cada agente estabelece relações positivas é um parâmetro do algoritmo e impacta diretamente na movimentação dos agentes no sistema.

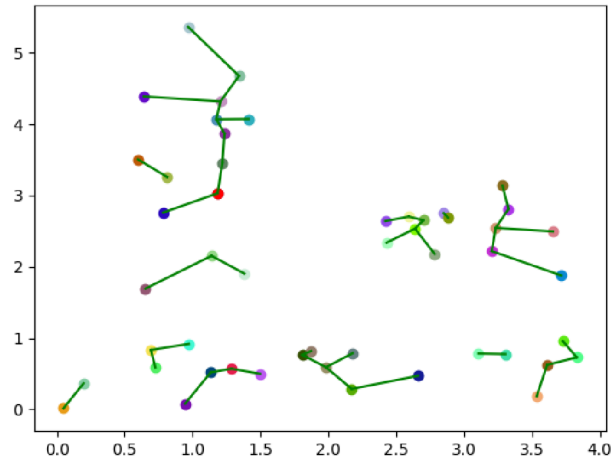
Na hipótese de ocorrerem conflitos de reciprocidade nestas relações, em eventuais situações em que determinado agente seja o mais próximo de outro e este não seja o mais próximo do primeiro, a relação entre os dois agentes é definida como positiva. As relações recíprocas, onde determinado agente é o mais próximo de outro e vice-versa ou determinado agente não é o mais próximo de outro e vice-versa, são apenas definidas respectivamente como positivas ou negativas.

A figura 4.4 apresenta um grafo de relações positivas, retratadas com arestas de cor verde entre os diversos agentes do sistema. No exemplo, a quantidade de agentes mais próximos com os quais cada agente estabelece relações positivas foi definido como 1. As relações negativas correspondem a todas as demais relações entre os agentes do sistema e foram suprimidas da ilustração em virtude da impossibilidade de distinção visual do total de 1225 arestas do grafo em um plano bidimensional.

4.3 Tomada de Decisão do Próximo Movimento pelos Agentes

Posteriormente, cada agente verifica o nível de conforto psicológico associado a todos setores do seu círculo de percepção no momento. Esta racionalização é feita por cálculos fundamentados na teoria da dissonância cognitiva de Festinger (1957) [20] e conceitos inerentes propostos por Heider (1958) [26] e Cartwright e Harary (1956) [10]. Estes conceitos são apresentados em

Figura 4.4: Grafo de relações positivas encontradas no ambiente.



detalhes na seção 2.

Na verificação, são destacadas apenas as relações interagentes dos indivíduos presentes em cada setor do círculo de percepção do agente em questão e calculado o produto das possíveis tríades destas relações, de forma a reproduzir o modelo de relações cognitivas conforme apresentado por Cartwright e Harary (1956) [10] para cada setor do círculo. Entende-se que para uma relação estar contida em um determinado setor, ambos agentes que a compõem devem estar presentes no determinado setor.

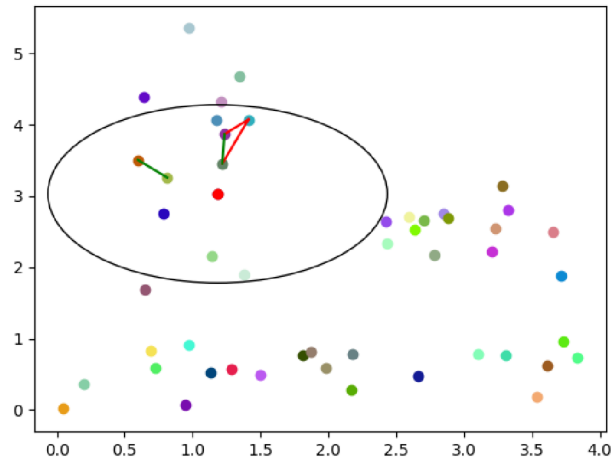
A figura 4.5 ilustra como um determinado agente percebe os conjuntos de relações presentes em cada setor do seu círculo de percepção. No exemplo, o agente observa que não há a presença de outros agentes nos setores 0 e 7 de seu círculo, bem como há apenas 1 agente presente nos setores 2, 4, 5 e 6, o que impossibilita a verificação de estado psicológico conforme modelo de relações de [26].

A fim de verificar todas as possibilidades de relações entre tríades de agentes dentre os identificadas em cada setor, cada relação positiva recebe o peso 1 e cada relação negativa recebe o peso -1. Conforme Heider (1958) [26], o produto do conjunto de relações que constituem uma tríade corresponde ao teor da mesma por inteiro. Isto possibilita que o teor psicológico de uma tríade de relações seja definido apenas como positivo ou negativo, respectivamente balanceado ou desbalanceado, confortável ou desconfortável.

Em conhecimento do teor psicológico de todas as possíveis tríades de relações presentes em cada setor, é calculado o teor psicológico do setor por inteiro. Conforme Flament [22], um grafo sinalizado completo, tal qual é apresentado para as verificações no presente algoritmo, pode ser classificado de acordo com seu grau de balanceamento. Outrossim, tal cálculo é utilizado para definição do estado psicológico de cada setor do círculo de percepção dos agentes do sistema, que posteriormente é resumido como positivo ou negativo, respectivamente, se o valor encontrado atingir o limiar de satisfação cognitiva, que é compartilhado pelos agentes do sistema.

A partir da quantidade total de relações interagentes presentes em uma situação, pode-

Figura 4.5: Detecção do teor psicológico por setores do círculo de percepção do agente em destaque.



se calcular a quantidade de tríades de relações interagentes presentes em uma situação. Da mesma forma, a partir da quantidade de relações positivas e negativas de uma situação, pode-se calcular a quantidade de tríades de relações interagentes de uma situação que são avaliadas como positivas. A área da matemática que descreve este tipo de cômputo é a análise combinatória [5].

A combinação simples é um método da análise combinatória que permite o cálculo da quantidade de possíveis subconjuntos do conjunto total, desconsiderando a ordenação dos elementos nos subconjuntos [5]. A equação 4.1 descreve esta operação, onde:

- r corresponde à quantidade de elementos no conjunto total; e
- k corresponde à quantidade de elementos nos subconjuntos cujas combinações serão calculadas.

$$C_{r,k} = \frac{r!}{k! \cdot (r - k)!} \quad (4.1)$$

Esta equação é aplicada no cálculo do balanceamento das diversas situações identificadas pelos agentes, especificamente para calcular o quantitativo de tríades de relações interagentes positivas a partir do quantitativo de relações positivas e negativas da respectiva situação ³. Enfim, calcula-se a soma das possíveis tríades de relações positivas e das possíveis tríades compostas por uma relação positiva e duas negativas. A equação 4.2 modela o cálculo realizado, onde:

- p corresponde à quantidade de relações positivas; e

³A subseção 2.5 lista as possíveis configurações de tríades de relações interagentes e como elas são avaliadas pela teoria do balanceamento psicológico, positivas ou negativas.

- n corresponde à quantidade de relações negativas.

A equação 4.2 é simplificada na forma de 4.3.

$$C_{p,n} = \frac{p!}{3! \cdot (p-3)!} + p \cdot \frac{n!}{2! \cdot (n-2)!} \quad (4.2)$$

$$C_{p,n} = \frac{p \cdot (p-1) \cdot (p-2)}{6} + \frac{p \cdot n \cdot (n-1)}{2} \quad (4.3)$$

Finalmente, o teor psicológico da situação é calculado na forma da equação 4.4, onde é determinada a razão entre o quantitativo de tríades positivas e o quantitativo de tríades positivas e negativas na situação analisada.

$$C_{p,n} = \frac{\frac{p \cdot (p-1) \cdot (p-2)}{6} + \frac{p \cdot n \cdot (n-1)}{2}}{\frac{(p+n) \cdot (p+n-1) \cdot (p+n-2)}{6}} \quad (4.4)$$

A figura 4.5 exemplifica, ainda, no setor 3 do círculo de percepção do agente em questão, a presença de apenas 2 outros indivíduos. Isto impossibilita o cálculo do teor psicológico do setor por meio da identificação de tríades de relações, visto que é evidenciada apenas uma relação. Contudo, Heider [26] sugere que um conjunto de relações composto por apenas duas entidades, sempre é de teor psicológico positivo.

Após conhecer o estado psicológico de cada setor de seu círculo de percepção, cada agente toma a decisão da direção do movimento a ser realizado na respectiva iteração. Esta decisão ocorre em função da direção do setor de conforto psicológico, de acordo com a análise do próprio agente. Outrossim, cada agente autonomamente escolhe por se mover em direção do centro do setor de maior conforto para si, tendo a dimensão de seu movimento pautada pela parametrização inicial do algoritmo.

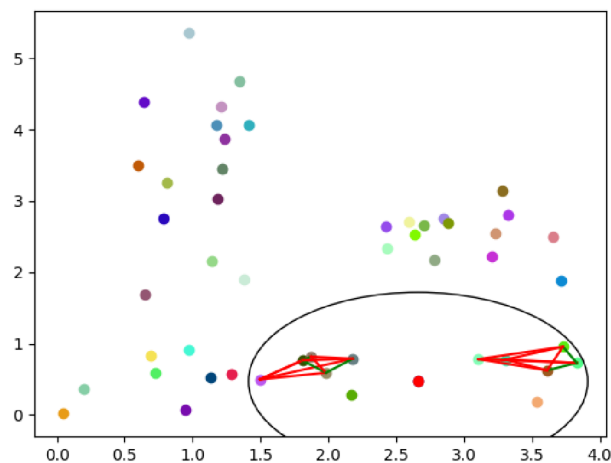
Ainda no exemplo da figura 4.5, é possível verificar que os setores 1 e 3 do círculo de percepção do agente em questão possuem agentes em quantidade suficiente para verificação de um subgrafo das relações entre eles, respectivamente 3 e 2 agentes. O setor 1 apresenta 1 relação positiva e 2 relações negativas, cujo produto resulta no balanço positivo para o referido subconjunto. O setor 3 apresenta apenas 1 relação, de caráter positivo, o que inerentemente resulta no balanço positivo para o referido subconjunto. Consequentemente, a escolha de movimento do agente em direção dos setores de seu círculo de percepção foi restrita àqueles cujos balanços cognitivos foram positivos.

Nas eventuais hipóteses de o agente encontrar mais do que um setor do seu círculo de percepção com o maior conforto psicológico ou não encontrar nenhum, hipótese na qual todos setores apresentariam estado psicológico de valor zero e, portanto, o maior conforto psicológico do setor, o agente escolherá, dentre estes, aquele de maior densidade populacional para direção de movimento correspondente a iteração.

O agente de comportamento retratado na figura 4.5 utiliza esta lógica em sua racionalização, visto que identifica dois de seus setores de percepção como de balanço cognitivo positivo, nominalmente os setores 1 e 3. Contudo, este opta pelo movimento em direção do setor 1, visto que este apresenta densidade de 3 agentes, em detrimento do setor 3, cuja densidade foi de apenas 2 agentes.

Também é possível que agentes não encontrem nenhum outro agente em seu campo de percepção, situação na qual eles não se movimentam na respectiva iteração em que o fenômeno ocorreu. Este comportamento é ilustrado na figura 4.6, onde é destacado o círculo de percepção de outro determinado agente do sistema. Neste caso, o agente encontrou os setores 1, 2, 5 e 6 do seu círculo de percepção sem a presença de outros agentes e os setores 4 e 7 com apenas 1 agente. Os setores 0 e 3 apresentaram, igualmente, 3 relações positivas em adição de 6 relações negativas, cujos produtos resultaram em balanço cognitivo negativo. Consequentemente, o agente destacado não realizou nenhum movimento na iteração retratada.

Figura 4.6: Exemplo de situação em que o agente em destaque não encontra setor do seu círculo de percepção com teor psicológico positivo, o que o impede de se movimentar na respectiva iteração.



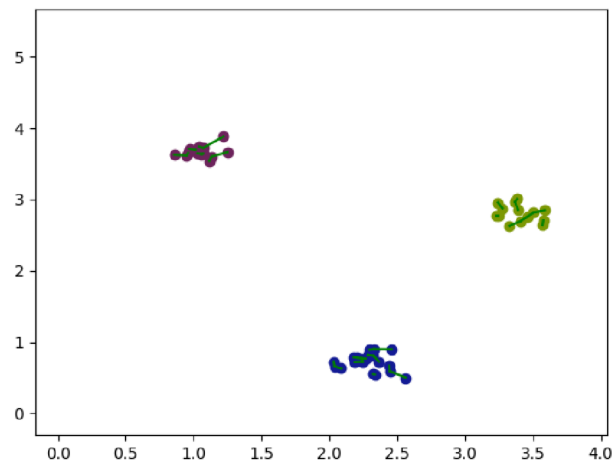
Ao tomar a decisão pela direção do movimento, que sempre vai ocorrer na direção do centro de um setor de seu círculo de percepção populado por outros agentes, e realizar o movimento no ambiente euclidiano, o agente também assume o agrupamento presente em tal setor que tenha maior representatividade em consideração do conjunto total de agentes do sistema. Caso ainda haja impasse na definição do agrupamento a ser assumido, o agente optará pelo agrupamento ao qual pertence o agente mais próximo de si e que pertença a um dos agrupamentos majoritários do setor escolhido para direção de movimento. Caso ainda persista o impasse, dos agentes em conflito, um é escolhido aleatoriamente. Desta forma, assegura-se a transferência de cognições e a identificação dos agrupamentos conforme os agentes se movem pelo ambiente.

A figura 4.5 mostra que o agente destacado em vermelho determina seu movimento da respectiva iteração em direção do setor 1, cujo teor psicológico foi determinado como positivo

e dentre os caracterizados desta forma, foi o de maior densidade. Ela também mostra que o agente presente neste setor e de maior proximidade com o agente colorido em vermelho é um colorido em verde, de forma que o agente em destaque assumirá o agrupamento e a cor deste até a próxima iteração, onde as características poderão ser modificadas novamente de acordo com sua percepção.

O processamento do algoritmo ocorre iterativamente, onde são recorrentes a distinção, por parte de cada agente, dos setores de seu círculo de percepção que proponham conforto psicológico em detrimento daqueles que proponham desconforto psicológico; daqueles setores de teor psicológico positivo, escolher os de maior densidade populacional e, na eventual igualdade de densidade entre setores, optar aleatoriamente entre estes; a realização do movimento de um passo em direção do centro do setor do círculo de percepção escolhido e; alteração do próprio agrupamento para participação no qual pertence o agente mais próximo de si dentro do setor escolhido. As iterações ocorrem até que nenhum agente do sistema mude de agrupamento em uma iteração completa ou que seja alcançada uma quantidade limite de iterações. A figura 4.7 ilustra a conformação final dos agentes do sistema, em termos de seus posicionamentos e como se aproximaram entre si, formando aglomerações conforme as iterações do algoritmo evoluíram.

Figura 4.7: Conformação final dos agentes do sistema.



Ao final do processamento, os agrupamentos aos quais pertencem cada agente do sistema podem ser traduzidos nos agrupamentos aos quais são atribuídos os pontos de dados do conjunto de entrada e este resultado corresponde a saída do algoritmo. A figura 4.8 apresenta a tradução do posicionamento final dos agentes no sistema em agrupamentos do conjunto de dados de entrada do algoritmo.

O algoritmo 1 apresenta o conjunto de instruções que definem o comportamento do Cognitive Consistency Clustering (C3). O algoritmo 2 descreve a verificação do teor psicológico de uma situação composta por um conjunto de relações interagentes.

Algorithm 1 Cognitive Consistency Clustering (C3)

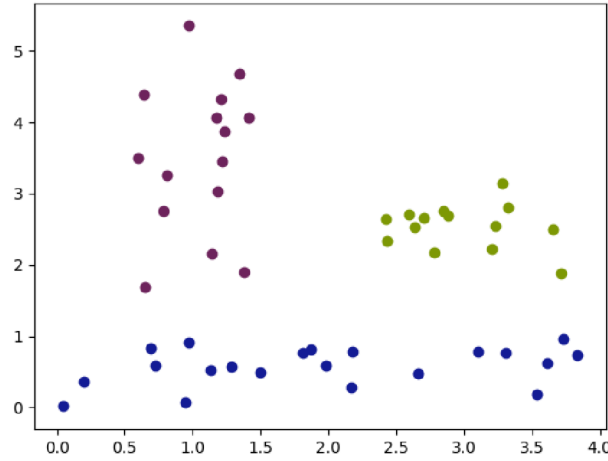
Entrada:

- 1► Um conjunto de dados $\Delta = \{ \delta_1, \delta_2, \dots, \delta_n \}$.
- 2► Extensão do raio de percepção dos agentes no sistema Γ .
- 3► Extensão do passo de movimentação dos agentes λ .
- 4► Quantidade de setores a dividir o círculo de percepção dos agentes ϕ .
- 5► Limiar de satisfação cognitiva ϵ .
- 6► Quantidade de agentes vizinhos mais próximos com os quais um agente estabelece relações positivas ϖ .

Saída:

- 7► Partições de Δ .
 - 8► **para todo** $\delta_i \in \Delta$ **faça**
 - 9► Instancie um agente α_i em A.
 - 10► Posicione α_i nas coordenadas euclidianas descritas em δ_i .
 - 11► Atribua a α_i um grupo κ único.
 - 12► **fim para**
 - 13► **enquanto** agrupamentos instáveis **faça** ▷ x
 - 14► $\Psi = \{ \psi_1, \psi_2, \dots, \psi_n \} \leftarrow$ conjunto de relações (L,L') entre $A = \{ \alpha_1, \alpha_2, \dots, \alpha_n \} : \varpi$.
 - 15► **para todo** $\alpha_i \in A$ **faça**
 - 16► Defina o círculo de percepção de α_i , θ , com raio Γ .
 - 17► Divida θ em ϕ setores, $\Omega = \{ \omega_1, \omega_2, \dots, \omega_n \}$.
 - 18► Em Ω , mantenha somente os ω de teor psicológico positivo, em consideração de Ψ .
(para verificação do teor psicológico de cada $\omega_i \in \Omega$, invoca-se o algoritmo 2)
 - 19► Em Ω , mantenha somente os ω cujas densidades sejam iguais a densidade do $\omega_i \in \Omega$ de maior densidade.
 - 20► $\tau \leftarrow$ escolha aleatoriamente um $\omega \in \Omega$.
 - 21► α_i se move λ unidades de distância na direção do centro de τ .
 - 22► α_i assume aleatoriamente um dos agrupamentos presentes em τ cujas densidades sejam iguais a daquele de maior densidade.
 - 23► **fim para**
 - 24► **fim enquanto**
 - 25► **para todo** $\delta_i \in \Delta$ **faça**
 - 26► $\delta_{i,\kappa} \leftarrow \alpha_{i,\kappa}$.
 - 27► **fim para**
-

Figura 4.8: Tradução da conformação final dos agentes do sistema em agrupamentos de dados.



Algorithm 2 Verificação do Teor Psicológico de um Setor de um Círculo de Percepção

Entrada:

- 1► Área de um setor de percepção ω de um agente $\alpha_i \in A = \{ \alpha_1, \alpha_2, \dots, \alpha_n \}$.
- 2► O conjunto de relações (L,L') $\Psi = \{ \psi_1, \psi_2, \dots, \psi_n \}$ entre $A = \{ \alpha_1, \alpha_2, \dots, \alpha_n \}$.
- 3► Limiar de satisfação cognitiva ϵ dos $\alpha_i \in A$.

Saída:

- 4► Teor psicológico ζ de ω , positivo ou negativo.
 - 5► $P = \{ \rho_1, \rho_2, \dots, \rho_n \} \subset \Psi$ | P estejam localizados totalmente dentro de ω .
 - 6► $N = \{ \nu_1, \nu_2, \dots, \nu_n \} \leftarrow$ possíveis tríades C_3^P de P .
 - 7► $M = \{ \mu_1, \mu_2, \dots, \mu_n \} \leftarrow$ produto das relações contidas em cada $\nu_i \in N$.
 - 8► $Z \leftarrow n (\mu_i \in M \mid \mu_i > 0) / n (\mu_i \in M)$.
 - 9► $\zeta = \begin{cases} \text{positivo} & \text{se } Z \geq \epsilon \\ \text{negativo} & \text{senão.} \end{cases}$
-

5 RESULTADOS OBTIDOS

Este capítulo descreve resultados obtidos pelo método de agrupamento de dados apresentado na seção 4. Estes resultados são apresentados em termos da comparação de desempenho entre o presente algoritmo, o Agent View Clustering (AVC) [29] e o K-means [44, 65] na detecção de agrupamentos de formatos e densidades distintos, em conjuntos de dados distintos ⁴.

O algoritmo AVC é particularmente apreciado no contexto da verificação de desempenho do algoritmo proposto no presente estudo devido a que parcialmente se assemelham nos processos de tomada de decisão sobre a direção de movimento dos agentes do sistema. Ambos utilizam o conceito de multiagentes na representação dos pontos de dados analisados e apresentam um processo iterativo onde eles decidem pela direção do movimento a ser realizado a partir da segmentação dos respectivos círculos de percepção e valorização destes segmentos em função da disposição de outros agentes nestas áreas.

Na comparação com a presente abordagem, a semelhança com AVC ocorre estritamente quando a escolha pelo movimento é pautada pelo critério de densidade. Enquanto, para este, a verificação de densidade nos setores do círculo de percepção dos agentes é o que pauta as respectivas tomadas de decisão para direção do movimento, sendo a escolha dada pelo setor de maior densidade, para a presente abordagem, o mesmo critério é apenas uma alternativa para a eventualidade de o critério principal, a verificação do balanço psicológico dos setores, não seja capaz de determinar um setor de balanço positivo para pautar o movimento. Nesta hipótese, dos setores do círculo de percepção dos agentes de balanço positivo, é escolhido aquele de maior densidade. Caso não haja setores de balanço positivo, não há movimento para o agente na iteração.

Cada algoritmo foi executado sobre o mesmo conjunto de dados no total de 10 ocorrências com diferentes estados iniciais de aleatoriedade. No total, foram utilizadas 9 métricas de avaliação de agrupamentos de dados sobre os 150 testes executados. Nominalmente, foram elas: os índices de Rand ajustado, informação mútua ajustada, homogeneidade, completude, métrica-V, Fowlkes-Mallows, Silhouette, Calinski-Harabasz e Davies-Bouldin.

Em termos de parametrização dos algoritmos utilizados nos testes, ambos o AVC e a presente abordagem tiveram seus hiperparâmetros de execução estimados pelo método de busca aleatória [4]. Inicialmente, foram realizadas tentativas de estimativa dos hiperparâmetros pelo método de busca exaustiva, mas estas não foram completas devido à quantidade de valores possíveis nos hiperparâmetros, cuja consequência foi explosão combinatória.

Não foram realizadas estimativas de hiperparâmetros para o algoritmo K-means devido a que as quantidades de agrupamentos presentes nos conjuntos de dados são conhecidas. Os hiperparâmetros utilizados nos três algoritmos aplicados são:

- k corresponde à quantidade de agrupamentos a serem descobertos;

⁴Os resultados parciais foram publicados em https://doi.org/10.1007/978-3-030-87897-9_17 [45]

- Γ corresponde ao raio de visão em que os agentes do sistema são capazes de detectar outros;
- λ corresponde ao tamanho da movimentação que o agente realiza em cada iteração do algoritmo;
- ϕ corresponde à quantidade de setores a dividir o círculo de percepção dos agentes;
- ϵ corresponde ao limiar de satisfação cognitiva que deve ser alcançado pelo conjunto de relações presentes em um setor do campo de visão de um agente para que seja considerado como opção viável para seu próximo movimento; e
- ϖ corresponde à quantidade de outros agentes mais próximos de si com os quais um agente estabelece relações positivas em uma iteração do algoritmo, enquanto estabelece relações negativas com todos demais.

5.1 Wine Dataset

O conjunto de dados de vinhos [23] é proveniente da análise química de vinhos de 3 variedades distintas, produzidos em uma região italiana. A análise determinou valores quantitativos para medição do volume de 13 substâncias presentes em 178 amostras. São elas o álcool, ácido málico, borralha, alcalinidade da borralha, magnésio, fenóis totais, flavonóides, fenóis não flavonóides, proantocianidinas, intensidade da cor, matiz, proteínas e prolina. Com exceção dos rótulos categóricos dos agrupamentos, os demais dados são numéricos.

Esta subseção apresenta resultados obtidos para tarefas de agrupamento realizadas com os algoritmos K-means, AVC e C3 sobre o conjunto de dados de vinhos. Os valores apresentados para as métricas de avaliação são referentes às médias de 10 execuções. Os hiperparâmetros utilizados neste conjunto de testes são relatados na tabela 1:

Tabela 1: Parâmetros utilizados nos testes sobre o conjunto de dados de vinhos.

Parâmetro	K-means	AVC	C3
k	3	-	-
Γ	-	1,8	1,8
λ	-	1,8	1,8
ϕ	-	8	8
ϵ	-	-	0,13
ϖ	-	-	39

Na avaliação pelo índice de Rand ajustado, os modelos descobertos pelo algoritmo K-means apresentaram desempenho médio de 0,87, um ganho de 27,9% sobre os modelos descobertos pelo algoritmo AVC, com desempenho médio de 0,68. Os modelos descobertos pelo algoritmo

C3 apresentaram Rand ajustado médio de 0,78, o que representa ganho de 14,7% sobre o desempenho médio dos modelos encontrados pelo algoritmo AVC, mas perda de 10,3% em comparação com os modelos encontrados pelo algoritmo K-means.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 18,8477 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre as médias de desempenho dos modelos descobertos pelo algoritmo C3 e pelos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 4,4392 e 4,2429, respectivamente.

Sob avaliação pelo índice de informação mútua, os modelos descobertos pelo algoritmo K-means apresentaram desempenho médio de 0,85, um ganho de 26,9% sobre os modelos descobertos pelo algoritmo AVC, com desempenho médio de 0,67. Os modelos descobertos pelo algoritmo C3 apresentaram informação mútua de 0,75, o que representa ganho de 11,9% sobre os modelos descobertos pelo algoritmo AVC, mas perda de 11,8% em comparação com os modelos descobertos pelo algoritmo K-means.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 25,9682 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre as três médias de desempenho encontradas são estatisticamente significantes. As estatísticas- Q , para as comparações de médias de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means foram, respectivamente, 4,4556 e 5,7104.

A avaliação de homogeneidade de agrupamentos apresentou valor médio de 0,86 para os modelos descobertos pelo algoritmo K-means, um ganho de 30,3% sobre os modelos descobertos pelo algoritmo AVC, com desempenho médio de 0,66. Os modelos descobertos pelo algoritmo C3 apresentaram homogeneidade média de 0,75, o que representa ganho de desempenho de 13,6% sobre os modelos descobertos pelo algoritmo AVC, mas perda de 12,8% em comparação com os modelos descobertos pelo algoritmo K-means.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 22,5435 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre as três médias de desempenho encontradas são estatisticamente significantes. As estatísticas- Q , para as comparações de médias de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means foram, respectivamente, 4,4270 e 5,0620.

A avaliação de completude de agrupamentos apresentou valor médio de 0,85 para os modelos descobertos pelo algoritmo K-means, um ganho de 23,2% sobre os modelos descobertos pelo algoritmo AVC, com desempenho médio de 0,69. Os modelos descobertos pelo algoritmo C3 apresentaram completude média de 0,75, o que representa ganho de desempenho de 8,7% sobre os modelos descobertos pelo algoritmo AVC, mas perda de desempenho de 11,8% em

comparação com os modelos descobertos pelo algoritmo K-means.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 31,5739 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{crítica}$ 3,5058 revelou que as diferenças entre as três médias de desempenho encontradas são estatisticamente significantes. As estatísticas- Q , para as comparações de médias de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means foram, respectivamente, 4,4573 e 6,7056.

Na avaliação pela métrica- V , os modelos descobertos pelo algoritmo K-means apresentaram valor médio de 0,86, o que representa ganho de 28,4% sobre os modelos descobertos pelo algoritmo AVC, com desempenho médio de 0,67. Os modelos descobertos pelo algoritmo C3 apresentaram métrica- V média de 0,75, o que representa ganho de 11,9% sobre os modelos descobertos pelo algoritmo AVC, mas perda de desempenho de 12,8% em comparação com os modelos descobertos pelo algoritmo K-means.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 26,0143 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{crítica}$ 3,5058 revelou que as diferenças entre as três médias de desempenho encontradas são estatisticamente significantes. As estatísticas- Q , para as comparações de médias de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means foram, respectivamente, 4,4561 e 5,7187.

Sob avaliação pelo índice de Fowlkes-Mallows, os modelos descobertos pelo algoritmo K-means apresentaram qualidade média de 0,91, o que representa ganho de 15,2% sobre o desempenho dos modelos descobertos pelo algoritmo AVC, com valor médio de 0,79. Os modelos descobertos pelo algoritmo C3 apresentaram Fowlkes-Mallows de 0,85, em média, o que representa ganho de desempenho de 7,6% sobre os modelos descobertos pelo algoritmo AVC, mas perda de desempenho de 6,6% em comparação com os modelos descobertos pelo algoritmo K-means.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 21,1201 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{crítica}$ 3,5058 revelou que as diferenças entre as três médias de desempenho encontradas são estatisticamente significantes. As estatísticas- Q , para as comparações de médias de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means foram, respectivamente, 4,4891 e 4,7014.

O índice de Silhouette para os modelos descobertos pelo algoritmo K-means foi de 0,56, em média, o que representa ganho de 33,3% sobre os modelos descobertos pelo algoritmo AVC, com valor médio de 0,42. Os modelos descobertos pelo algoritmo C3 apresentaram Silhouette médio de 0,51, o que representa ganho de desempenho de 21,4% sobre os modelos descobertos pelo algoritmo AVC, mas perda de desempenho de 8,9% em comparação com os modelos descobertos

pelo algoritmo K-means.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 11,8197 e valor- p de 0,0002, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{crítica}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmo C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes. As estatísticas- Q , para as comparações de médias de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means foram, respectivamente, 4,3361 e 2,4534.

Sob avaliação pelo índice de Calinski-Harabasz, os modelos descobertos pelo algoritmo K-means apresentaram valor médio de 343,83, o que representa ganho de 76,4% sobre o desempenho médio dos modelos descobertos pelo algoritmo AVC, com valor de 194,94. Os modelos descobertos pelo algoritmo C3 apresentaram Calinski-Harabasz de 272,62, indicando ganho de desempenho de 39,8% sobre os modelos descobertos pelo algoritmo AVC, mas perda de desempenho de 20,7% na comparação com os modelos descobertos pelo algoritmo K-means.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 17,9669 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{crítica}$ 3,5058 revelou que as diferenças entre as três médias de desempenho encontradas são estatisticamente significantes. As estatísticas- Q , para as comparações de médias de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means foram, respectivamente, 4,4216 e 4,0532.

O índice de Davies-Bouldin para os modelos descobertos pelo algoritmo K-means foi de 0,60, em média, o que representa ganho de 27,7% sobre os modelos descobertos pelo algoritmo AVC, com valor de 0,83. Os modelos descobertos pelo algoritmo C3 apresentaram Davies-Bouldin médio de 0,67, o que representa ganho de desempenho de 19,3% sobre os modelos descobertos pelo algoritmo AVC, mas perda de desempenho de 11,7% na comparação com os modelos descobertos pelo algoritmo K-means.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 10,1416 e valor- p de 0,0005, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{crítica}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmo C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes. As estatísticas- Q , para as comparações de médias de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means foram, respectivamente, 4,4216 e 4,0532.

Todos os testes de Tukey para diferença honestamente significativa entre as médias de desempenho dos algoritmos C3, AVC e K-means, sob 9 métricas de qualidade de agrupamentos de dados, 6 internas e 3 externas, corroboraram com a superioridade de desempenho do algoritmo C3 sobre o algoritmo AVC, mas não sobre o algoritmo K-means.

A figura 5.1 apresenta uma comparação visual entre os resultados obtidos pelos modelos obtidos em média pelos 3 algoritmos utilizados nas métricas utilizadas na verificação de desem-

penho apresentada neste capítulo. Na figura, são destacadas as médias de resultados obtidos por algoritmo e métrica. A tabela 2 apresenta os dados utilizados na formatação dos gráficos presentes na figura 5.1 em formato tabular.

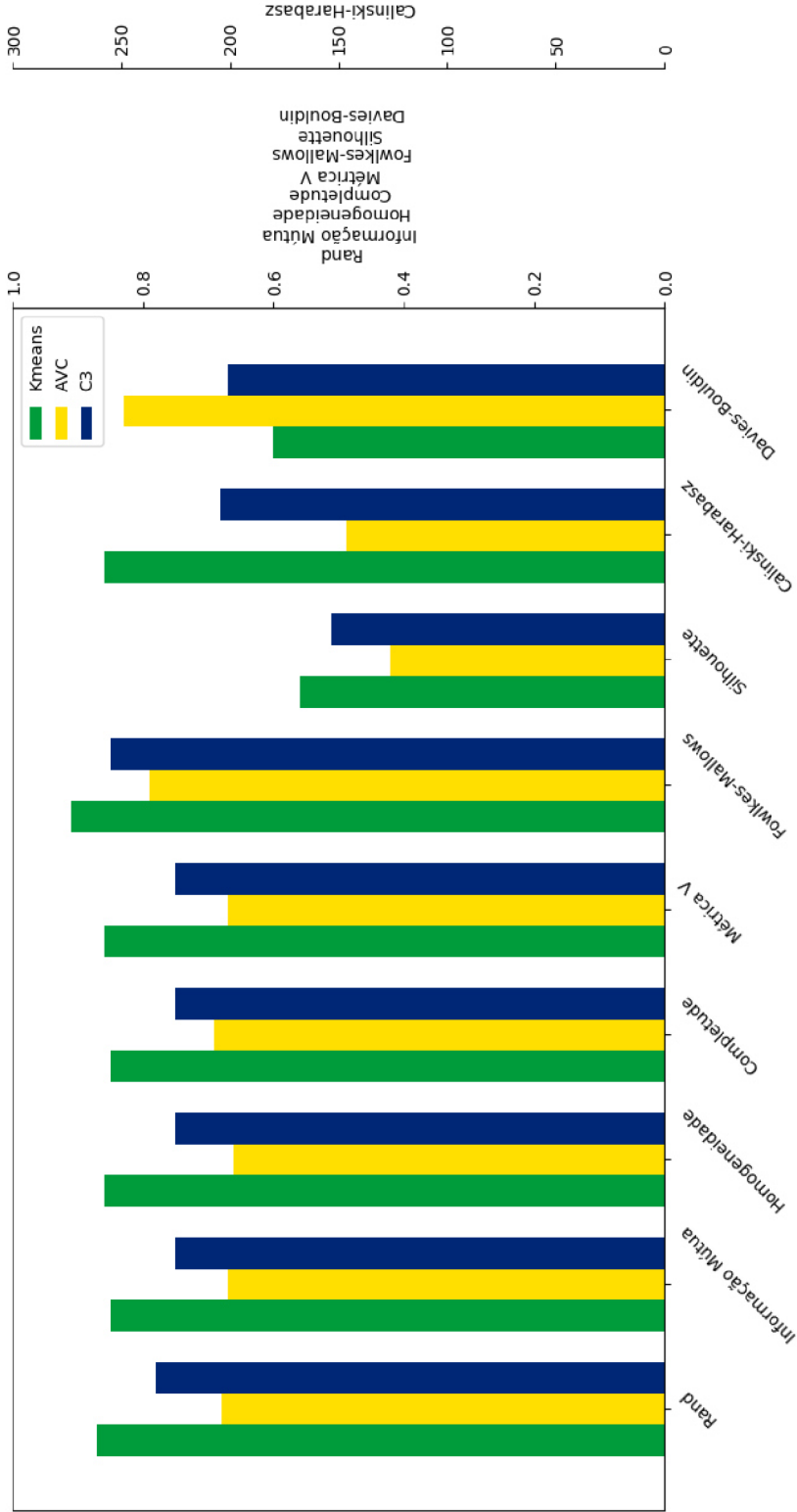
A figura 5.2 exhibe a conformação original dos agrupamentos (a), a conformação dos agrupamentos realizados pelo algoritmo K-means (b), a conformação dos agrupamentos realizados pelo algoritmo AVC (c) e a conformação dos agrupamentos realizados pelo algoritmo C3.

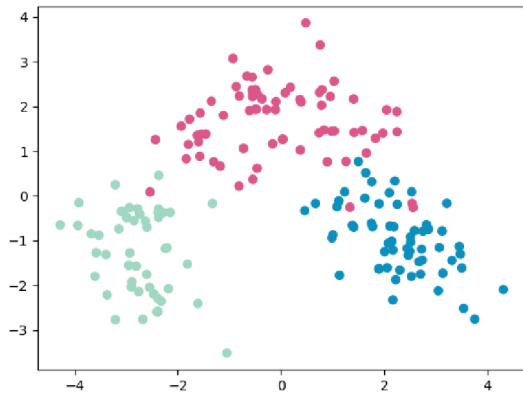
Na conformação original dos agrupamentos do conjunto de dados, foram calculados os graus de definição 0,53, 297,98 e 0,64, respectivamente, pelos índices internos para avaliação de agrupamentos 2 Silhouette, Calinski-Harabasz e Davies-Bouldin. Estas informações também são relacionadas à boa ou má definição das fronteiras dos agrupamentos originais do conjunto de dados analisado.

Tabela 2: Indicadores de qualidade de agrupamentos de dados (Wine).

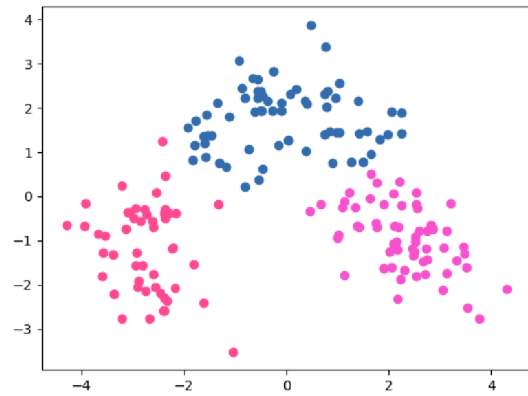
Algoritmo	Rand	Info Mútua	Homogeneidade	Compleitude	V	Fowlkes-Mallows	Silhouette	Calinski-Harabasz	Davies-Bouldin
K-means	0,87	0,85	0,86	0,85	0,86	0,91	0,56	343,83	0,60
AVC	0,68	0,67	0,66	0,69	0,67	0,79	0,42	194,94	0,83
C3	0,78	0,75	0,75	0,75	0,75	0,85	0,51	272,62	0,67

Figura 5.1: Comparativo visual de desempenho dos algoritmos K-means, AVC e C3 (Wine).

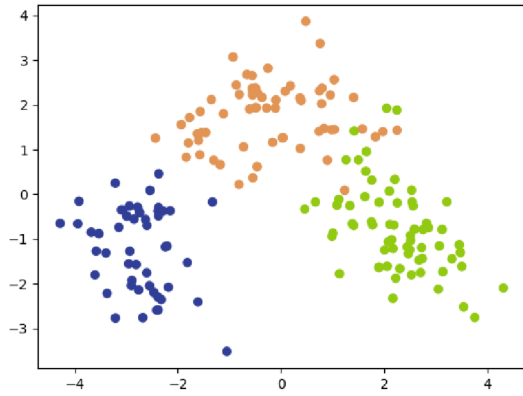




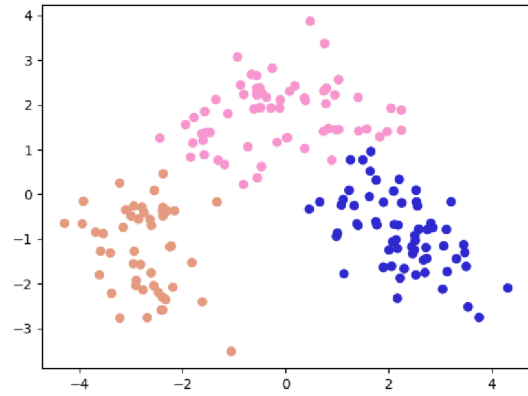
(a) Agrupamentos Originais



(b) Agrupamentos por K-means



(c) Agrupamentos por AVC



(d) Agrupamentos por C3

Figura 5.2

5.2 Zoo Dataset

O conjunto de dados de animais do zoológico é constituído de 17 atributos, referentes às características de 101 animais [24]. Nominalmente, são eles: presenças de pêlos, penas, denticção, veneno, barbatanas, cauda e coluna vertebral; capacidades de voar, nadar, respirar, predatória e amamentação; físico de grande porte; oviparidade; domesticação; quantidade de pernas e a classe ao qual pertencem. As possíveis classes são mamíferos, pássaros, répteis, peixes, anfíbios, insetos e invertebrados. Com exceção dos rótulos categóricos dos agrupamentos, os demais dados são numéricos.

Esta subseção apresenta resultados obtidos para tarefas de agrupamento realizadas com os algoritmos K-means, AVC e C3 sobre o conjunto de dados de animais do zoológico. Os valores apresentados para as métricas de avaliação são referentes às médias de 10 execuções consecutivas. Os hiperparâmetros utilizados neste conjunto de testes são relatados na tabela 3:

Na avaliação pelo índice de Rand ajustado, os modelos descobertos pelo algoritmo AVC apresentaram desempenho médio de 0,56, um ganho de 14,3% sobre os modelos descobertos pelo

Tabela 3: Parâmetros utilizados nos testes sobre o conjunto de dados de animais do zoológico.

Parâmetro	K-means	AVC	C3
k	7	-	-
Γ	-	1,3	1,3
λ	-	1,3	1,3
ϕ	-	8	8
ϵ	-	-	0,15
ϖ	-	-	20

algoritmo K-means, com desempenho médio de 0,49. Os modelos descobertos pelo algoritmo C3 apresentaram Rand ajustado de 0,68, o que representa ganhos de 38,8% e 21,4% sobre os modelos apresentados pelos algoritmos K-means e AVC, respectivamente.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 9,7477 e valor- p de 0,0007, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmo C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes. As estatísticas- Q , para as comparações de médias de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means foram, respectivamente, 3,9503 e 6,1631.

Sob avaliação pelo índice de informação mútua, os modelos descobertos pelo algoritmo K-means apresentaram desempenho médio de 0,63, um ganho de 5,0% sobre os modelos descobertos pelo algoritmo AVC, com desempenho de 0,60. Os modelos descobertos pelo algoritmo C3 apresentaram informação mútua de 0,56, o que representa perdas de desempenho de 11,1% e 6,7% em comparação com os modelos descobertos pelos algoritmos K-means e AVC, respectivamente.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 8,1319 e valor- p de 0,0017, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que a diferença entre as médias de desempenho dos algoritmos C3 e K-means são estatisticamente significantes. As estatísticas- Q , para as comparações de médias de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means foram, respectivamente, 3,1132 e 5,6950.

A avaliação de homogeneidade de agrupamentos apresentou valor médio de 0,70 para os modelos descobertos pelo algoritmo K-means, um ganho de 37,3% sobre o modelo descoberto pelo algoritmo AVC, com desempenho de 0,51. Os modelos descobertos pelo algoritmo C3 apresentaram homogeneidade média de 0,69, o que representa ganho de desempenho de 35,3% sobre os modelos descobertos pelo algoritmo AVC, mas perda de desempenho de 1,4% em comparação com os modelos descobertos pelo algoritmo K-means.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 72,2986 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de

desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que a diferença entre as médias de desempenho dos algoritmos C3 e AVC são estatisticamente significantes. As estatísticas- Q , para as comparações de médias de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means foram, respectivamente, 14,3871 e 0,6585.

A avaliação de completude de agrupamentos apresentou valor médio de 0,80 para os modelos descobertos pelo algoritmo AVC, um ganho de 23,1% sobre os modelos descobertos pelo algoritmo K-means, com desempenho médio de 0,65. Os modelos descobertos pelo algoritmo C3 apresentaram completude média de 0,61, o que representa perdas de desempenho de 6,2% e 23,8% em comparação com os modelos descobertos pelos algoritmos K-means e AVC, respectivamente.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 65,5625 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre as médias de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes. As estatísticas- Q , para as comparações de médias de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means foram, respectivamente, 15,4767 e 3,6102.

Na avaliação pela métrica-V, os modelos descobertos pelo algoritmo K-means apresentaram valor médio de 0,67, o que representa ganho de desempenho de 8,1% sobre os modelos descobertos pelo algoritmo AVC, com desempenho de 0,62. Os modelos descobertos pelo algoritmo C3 apresentaram métrica-V de 0,64, o que representa ganho de 3,2% de desempenho sobre os modelos descobertos pelo algoritmo AVC, mas perda de 4,5% de desempenho sobre os modelos descobertos pelo algoritmo K-means.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 5,3838 e valor- p de 0,0108, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre as médias de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente insignificantes. As estatísticas- Q , para as comparações de médias de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means foram, respectivamente, 1,8976 e 2,7188.

Sob avaliação pelo índice de Fowlkes-Mallows, os modelos descobertos pelo algoritmo AVC apresentaram qualidade média de 0,70, o que representa ganho de 16,7% sobre o desempenho do modelo descoberto pelo algoritmo K-means, com valor de 0,60. Os modelos descobertos pelo algoritmo C3 apresentaram Fowlkes-Mallows médio de 0,76, o que representa ganhos de 26,7% e 8,6% sobre os modelos apresentados pelos algoritmos K-means e AVC, respectivamente.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 10,6167 e valor- p de 0,0004, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que a diferença entre a média de desempenho do

algoritmos C3 e a média de desempenho do algoritmo K-means são estatisticamente significantes. As estatísticas- Q , para as comparações de médias de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means foram, respectivamente, 2,1940 e 6,4111.

O índice de Silhouette para os modelos descobertos pelo algoritmo AVC foi de 0,53, em média, o que representa ganhos de desempenho de 1,9% e 430,0% sobre os modelos descobertos pelos algoritmos K-means e C3, respectivamente. Os modelos descobertos pelo algoritmo C3 apresentaram Silhouette médio de 0,10, o que representou perdas de desempenho de 81,0% em comparação com os modelos descobertos pelos algoritmo K-means e AVC. Os modelos descobertos pelo algoritmo K-means apresentaram Silhouette médio de 0,52.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 167,1702 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{crítica}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes. As estatísticas- Q , para as comparações de médias de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means foram, respectivamente, 22,6842 e 22,0930.

Sob avaliação pelo índice de Calinski-Harabasz, os modelos descobertos pelo algoritmo K-means apresentaram valor médio de 244,67, o que representa ganhos de 93,0% e 679,2% sobre os desempenhos dos modelos descobertos pelos algoritmos AVC e C3, respectivamente. O algoritmo C3 apresentou modelos com Calinski-Harabasz de 31,40, enquanto o algoritmo AVC apresentou modelos de 126,78 para a mesma métrica, o que representa ganho de desempenho de 303,8% em favor deste.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 67,3325 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{crítica}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes. As estatísticas- Q , para as comparações de médias de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means foram, respectivamente, 7,3259 e 16,3809.

O índice de Davies-Bouldin para os modelos descobertos pelo algoritmo AVC foi de 0,59, em média, o que representa ganho de 16,9% sobre os modelos descobertos pelo algoritmo K-means, com valor de 0,71. Os modelos descobertos pelo algoritmo C3 apresentaram desempenho médio de 3,38 pelo índice de Davies-Bouldin, o que representa perdas de desempenho de 376,1% e 472,9% na comparação com os modelos descobertos pelos algoritmos K-means e AVC, respectivamente.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 31,5085 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{crítica}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatistica-

mente significantes. As estatísticas- Q , para as comparações de médias de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means foram, respectivamente, 9,9270 e 9,5041.

Os testes de Tukey para diferença honestamente significativa entre as médias de desempenho dos algoritmos C3, AVC e K-means corroboraram com a superioridade do algoritmo C3 sobre o AVC sob as métricas de qualidade de agrupamentos internas Rand ajustado e homogeneidade. A superioridade da abordagem proposta sobre o algoritmo K-means foi corroborada pelas métricas Rand ajustado e Fowlkes-Mallows.

A figura 5.3 apresenta uma comparação visual entre os resultados obtidos pelos modelos obtidos em média pelos 3 algoritmos utilizados nas métricas utilizadas na verificação de desempenho apresentada neste capítulo. Na figura, são destacadas as médias de resultados obtidos por algoritmo e métrica. A tabela 4 apresenta os dados utilizados na formatação dos gráficos presentes na figura 5.3 em formato tabular.

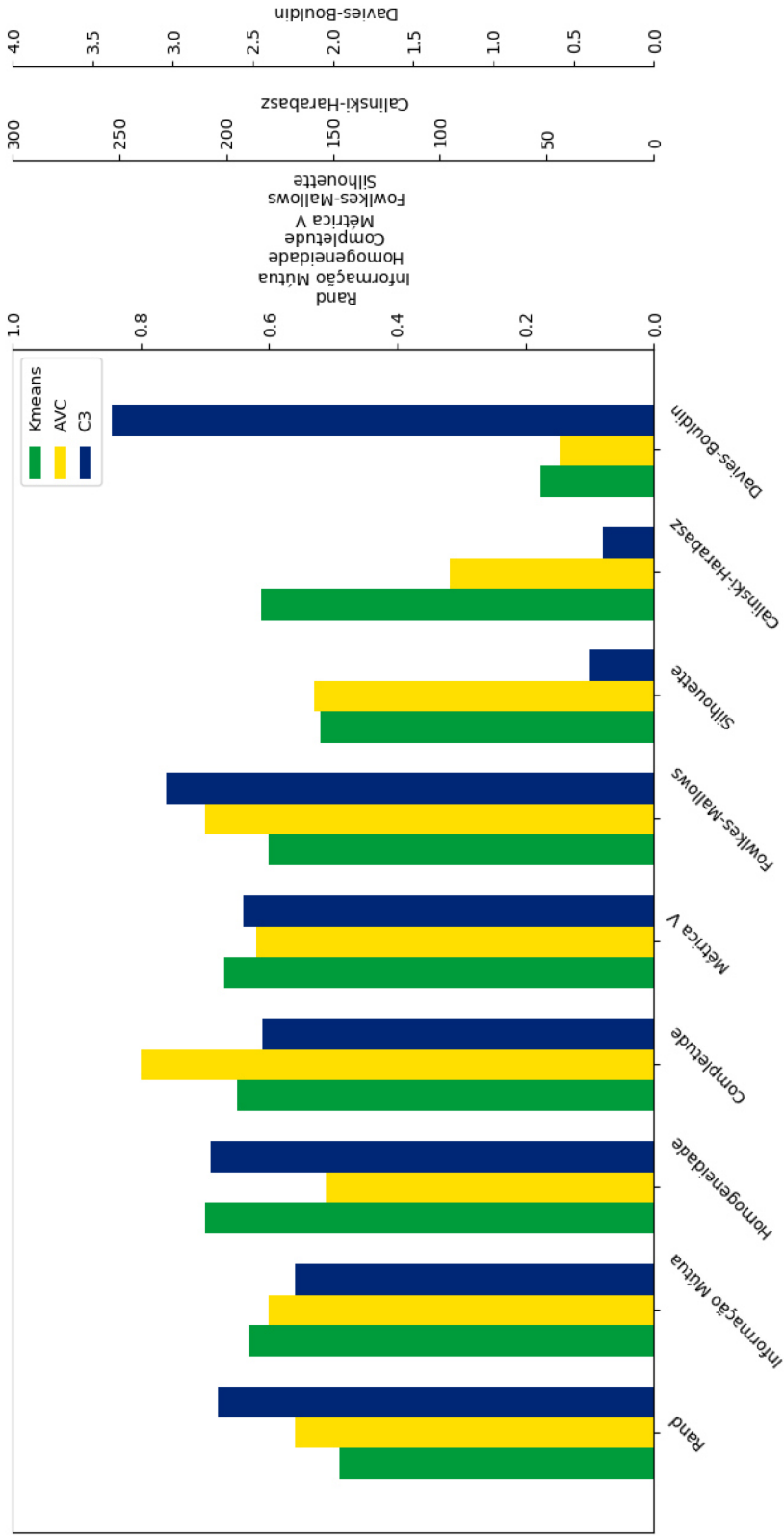
A figura 5.4 exhibe a conformação original dos agrupamentos (a), a conformação dos agrupamentos realizados pelo algoritmo K-means (b), a conformação dos agrupamentos realizados pelo algoritmo AVC (c) e a conformação dos agrupamentos realizados pelo algoritmo C3.

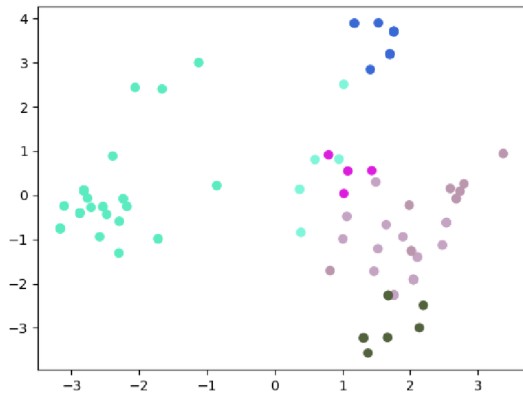
Na conformação original dos agrupamentos do conjunto de dados, foram calculados os graus de definição 0,42, 127,33 e 1,46, respectivamente, pelos índices internos para avaliação de agrupamentos 2 Silhouette, Calinski-Harabasz e Davies-Bouldin. Estas informações também são relacionadas à boa ou má definição das fronteiras dos agrupamentos originais do conjunto de dados analisado.

Tabela 4: Indicadores de qualidade de agrupamentos de dados (Zoo).

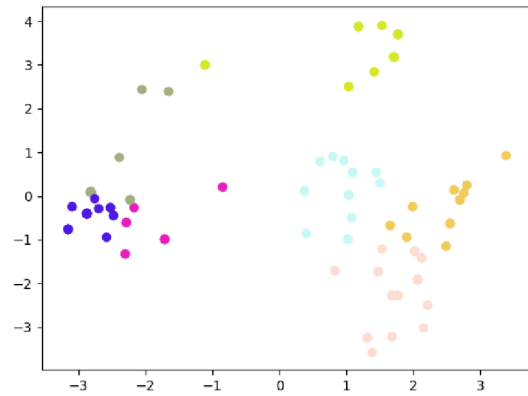
Algoritmo	Rand	Info Mútua	Homogeneidade	Compleitude	V	Fowlkes-Mallows	Silhouette	Calinski-Harabasz	Davies-Bouldin
K-means	0,49	0,63	0,70	0,65	0,67	0,60	0,52	244,67	0,71
AVC	0,56	0,60	0,51	0,80	0,62	0,70	0,53	126,78	0,59
C3	0,68	0,56	0,69	0,61	0,64	0,76	0,10	31,40	3,38

Figura 5.3: Comparativo visual de desempenho dos algoritmos K-means, AVC e C3 (Zoo).

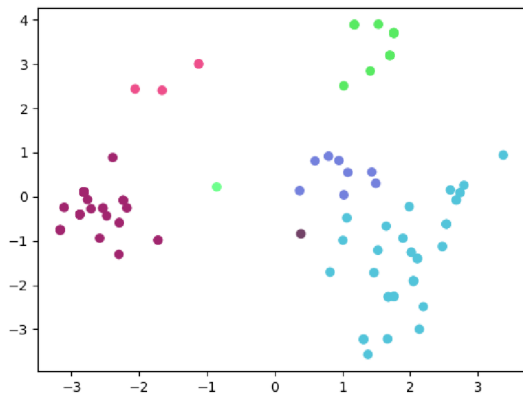




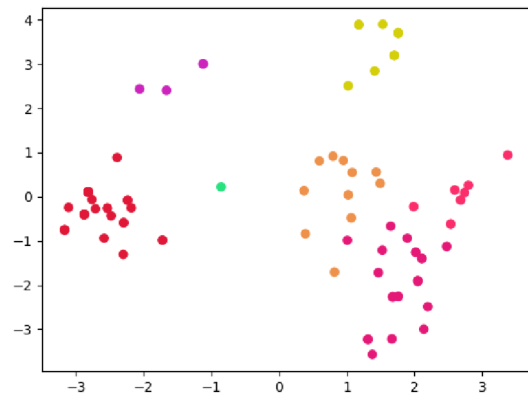
(a) Agrupamentos Originais



(b) Agrupamentos por K-means



(c) Agrupamentos por AVC



(d) Agrupamentos por C3

Figura 5.4

5.3 Thyroid Dataset

O conjunto de dados thyroid [14] foi criado para estudar a previsão de distúrbios da tireóide em humanos, onde os agrupamentos originais foram rotulados como hipotireoidismo, hipertireoidismo e saudável. As variáveis presentes no conjunto de dados são referentes a anamnese e resultados de exames nos 215 pacientes participantes do estudo. Foram elas o percentual de absorção de resina T3, tiroxina sérica total medida pelo método de deslocamento isotópico, triiodotironina sérica total medida por radioimuniensaio, hormônio estimulante da tireóide basal conforme medido por radioimuniensaio e diferença absoluta máxima do valor do hormônio tireoestimulante após injeção de 200 microgramas de hormônio liberador de tireotropina em comparação com o valor basal. Com exceção dos rótulos categóricos dos agrupamentos, os demais dados são numéricos.

Esta subseção apresenta resultados obtidos para tarefas de agrupamento realizadas com os algoritmos K-means, AVC e C3 sobre o conjunto de dados thyroid. Os valores apresentados para as métricas de avaliação são referentes às médias de 10 execuções consecutivas. Os

hiperparâmetros utilizados neste conjunto de testes são relatados na tabela 5:

Tabela 5: Parâmetros utilizados nos testes sobre o conjunto de dados thyroid.

Parâmetro	K-means	AVC	C3
k	3	-	-
Γ	-	0,5	0,5
λ	-	0,5	0,5
ϕ	-	8	8
ϵ	-	-	0,05
ϖ	-	-	6

Na avaliação pelo índice de Rand ajustado, os modelos descobertos pelo algoritmo AVC apresentaram desempenho médio de 0,70, um ganho de 18,6% sobre os modelos descobertos pelo algoritmo K-means, com desempenho médio de 0,59. Os modelos descobertos pelo algoritmo C3 apresentaram Rand ajustado de 0,72, o que representa ganhos de 22,0% e 2,9% sobre os modelos apresentados pelos algoritmos K-means e AVC, respectivamente.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 595,8774 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 9,2537 e 46,1408, respectivamente.

Sob avaliação pelo índice de informação mútua, os modelos descobertos pelo algoritmo AVC apresentaram desempenho médio de 0,48, uma perda de 14,3% sobre os modelos descobertos pelo algoritmo K-means, com desempenho médio de 0,56. Os modelos descobertos pelo algoritmo C3 apresentaram informação mútua de 0,51, em média, o que representa ganho de 6,3% sobre os modelos apresentados pelo algoritmo AVC, mas com perda de desempenho de 8,9% em comparação com os modelos apresentados pelo algoritmo K-means.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 118,1483 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 7,2120 e 14,1545, respectivamente.

A avaliação de homogeneidade de agrupamentos apresentou valor médio de 0,64 para os modelos descobertos pelo algoritmo AVC, um ganho de 30,6% sobre os modelos descobertos pelo algoritmo K-means, com desempenho médio de 0,49. Os modelos descobertos pelo algoritmo C3 apresentaram homogeneidade de 0,66, o que representa ganhos de 34,7% e 3,1% sobre os modelos apresentados pelos algoritmos K-means e AVC, respectivamente.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 1.259,0913 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de

desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 7,6221 e 64,9151, respectivamente.

A avaliação de completude de agrupamentos apresentou valor médio de 0,67 para os modelos descobertos pelo algoritmo K-means, o que representou superioridade de desempenho em comparação com os modelos descobertos pelos algoritmos AVC e C3 de 48,9% e 42,6%, respectivamente. Os modelos descobertos pelo algoritmo C3 apresentaram completude média de 0,47, o que representa ganho de 4,4% sobre os modelos descobertos pelo algoritmo AVC, com valor médio de 0,45 para a mesma métrica.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 1.097,0708 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 5,6047 e 54,3611, respectivamente.

Na avaliação pela métrica-V, os modelos descobertos pelo algoritmo K-means apresentou valor médio de 0,57, o que representou ganho de 7,5% sobre o modelo descoberto pelo algoritmo AVC, com desempenho de 0,53. Os modelos descobertos pelo algoritmo C3 apresentaram métrica-V 0,55 em média, o que representa ganho de 3,8% de desempenho sobre os modelos descobertos pelo algoritmo AVC, mas perda de 3,5% de desempenho sobre os modelos descobertos pelo algoritmo K-means.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 43,6653 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 8,0084 e 5,1004, respectivamente.

Sob avaliação pelo índice de Fowlkes-Mallows, os modelos descobertos pelo algoritmo AVC apresentaram qualidade média de 0,87, o que representa ganho de 3,6% sobre os desempenhos dos modelos descobertos pelo algoritmo K-means, com valor médio de 0,84. Os modelos descobertos pelo algoritmo C3 apresentaram Fowlkes-Mallows de 0,88, o que representa ganhos de 4,8% e 1,1% sobre os modelos apresentados pelos algoritmos K-means e AVC, respectivamente.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 282,1464 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 9,9671 e 32,7672, respectivamente.

O índice de Silhouette para os modelos descobertos pelo algoritmo K-means foi de 0,73, em

média, o que representa ganho de 28,1% sobre os modelos descobertos pelo algoritmo AVC, com valor médio de 0,32 para a mesma métrica. Os modelos descobertos pelo algoritmo C3 apresentou Silhouette de 0,38, o que representou ganho de desempenho de 18,8% sobre os modelos descobertos pelo algoritmo AVC, mas perda de 47,9% sobre os modelos descobertos pelo algoritmo K-means.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 245,2263 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 4,1836 e 24,7886, respectivamente.

Sob avaliação pelo índice de Calinski-Harabasz, os modelos descobertos pelo algoritmo K-means apresentaram valor médio de 345,86, o que representa ganhos de desempenho de 242,9% e 197,2% sobre os modelos descobertos pelos algoritmos AVC e C3, respectivamente. O algoritmo C3 apresentou modelos com Calinski-Harabasz de 116,37, enquanto o algoritmo AVC apresentou modelos de 100,85 para a mesma métrica, o que representa ganho de desempenho de 15,4% em favor do primeiro.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 2.813,4244 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 5,9992 e 88,7244, respectivamente.

O índice de Davies-Bouldin para os modelos descobertos pelo algoritmo K-means foi de 0,51, em média, o que representa ganho de 20,3% sobre os modelos descobertos pelo algoritmo AVC, com valor médio de 0,64. Os modelos descobertos pelo algoritmo C3 apresentou desempenho de 0,46 pelo índice de Davies-Bouldin, o que representa ganhos de desempenho de 9,8% e 28,1% em comparação com os modelos descobertos pelo algoritmo K-means e AVC, respectivamente.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 1,0492 e valor- p de 0,3641, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente insignificantes, com estatísticas- Q de 1,9669 e 0,4873, respectivamente.

Os testes de Tukey para diferença honestamente significativa entre as médias de desempenho dos algoritmos C3, AVC e K-means corroboraram com a superioridade do algoritmo C3 sobre o AVC sob as métricas de qualidade de agrupamentos internas Rand ajustado, informação mútua ajustada, homogeneidade, completude, métrica-V e Fowlkes-Mallows; e externas Silhouette e Calinski-Harabasz. A superioridade da abordagem proposta sobre o algoritmo K-means foi corroborada por testes de Tukey sobre as médias de desempenho sob as métricas internas Rand

ajustado, homogeneidade e Fowlkes-Mallows.

A figura 5.5 apresenta uma comparação visual entre os resultados obtidos pelos modelos obtidos em média pelos 3 algoritmos utilizados nas métricas utilizadas na verificação de desempenho apresentada neste capítulo. Na figura, são destacadas as médias de resultados obtidos por algoritmo e métrica. A tabela 6 apresenta os dados utilizados na formatação dos gráficos presentes na figura 5.5 em formato tabular.

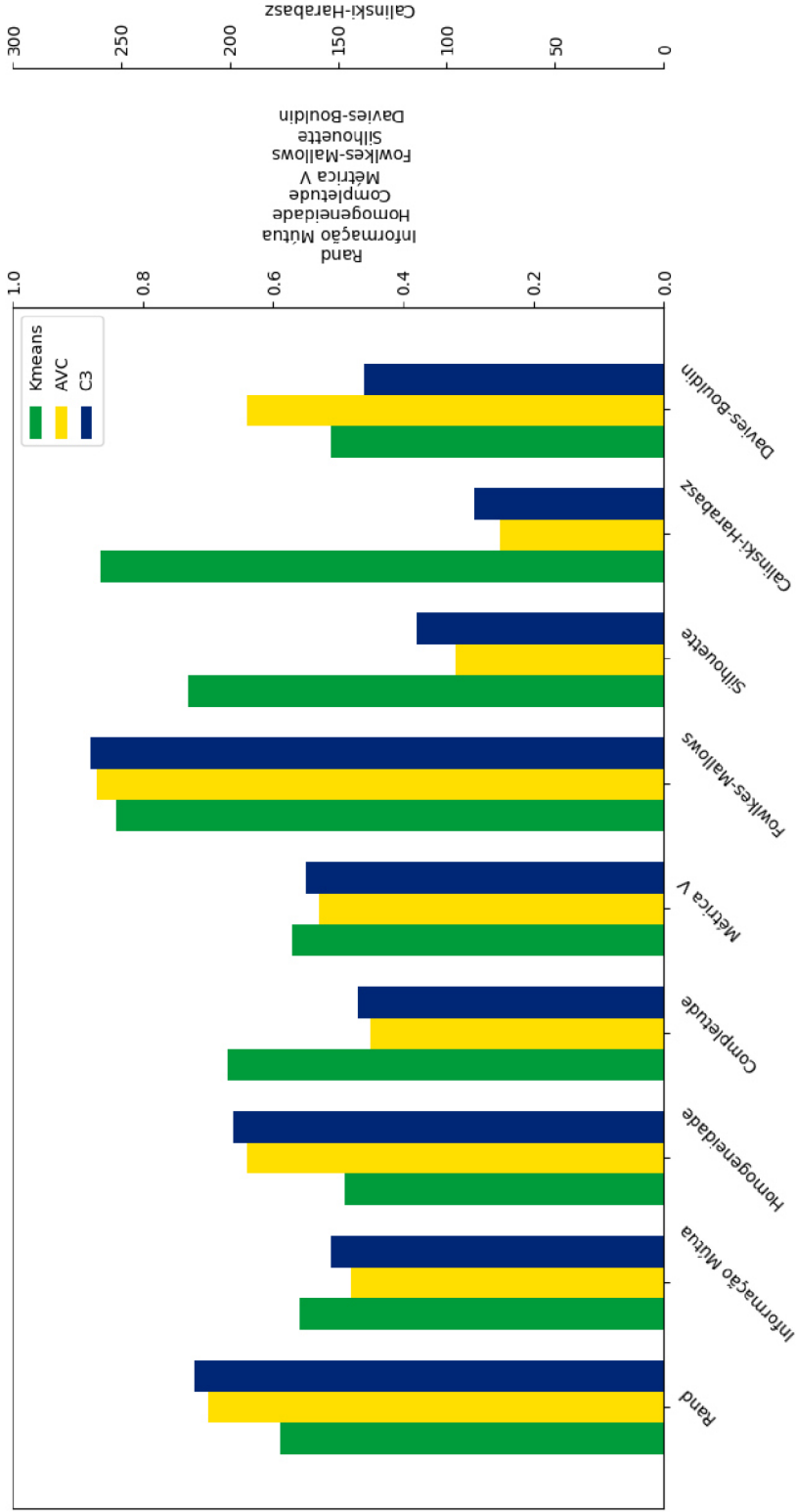
A figura 5.6 exibe a conformação original dos agrupamentos (a), a conformação dos agrupamentos realizados pelo algoritmo K-means (b), a conformação dos agrupamentos realizados pelo algoritmo AVC (c) e a conformação dos agrupamentos realizados pelo algoritmo C3.

Na conformação original dos agrupamentos do conjunto de dados, foram calculados os graus de definição 0,59, 195,93 e 0,71, respectivamente, pelos índices internos para avaliação de agrupamentos 2 Silhouette, Calinski-Harabasz e Davies-Bouldin. Estas informações também são relacionadas à boa ou má definição das fronteiras dos agrupamentos originais do conjunto de dados analisado.

Tabela 6: Indicadores de qualidade de agrupamentos de dados (Thyroid).

Algoritmo	Rand	Info Mútua	Homogeneidade	Compleitude	V	Fowlkes-Mallows	Silhouette	Calinski-Harabasz	Davies-Bouldin
K-means	0,59	0,56	0,49	0,67	0,57	0,84	0,73	345,86	0,51
AVC	0,70	0,48	0,64	0,45	0,53	0,87	0,32	100,85	0,64
C3	0,72	0,51	0,66	0,47	0,55	0,88	0,38	116,37	0,46

Figura 5.5: Comparativo visual de desempenho dos algoritmos K-means, AVC e C3 (Thyroid).



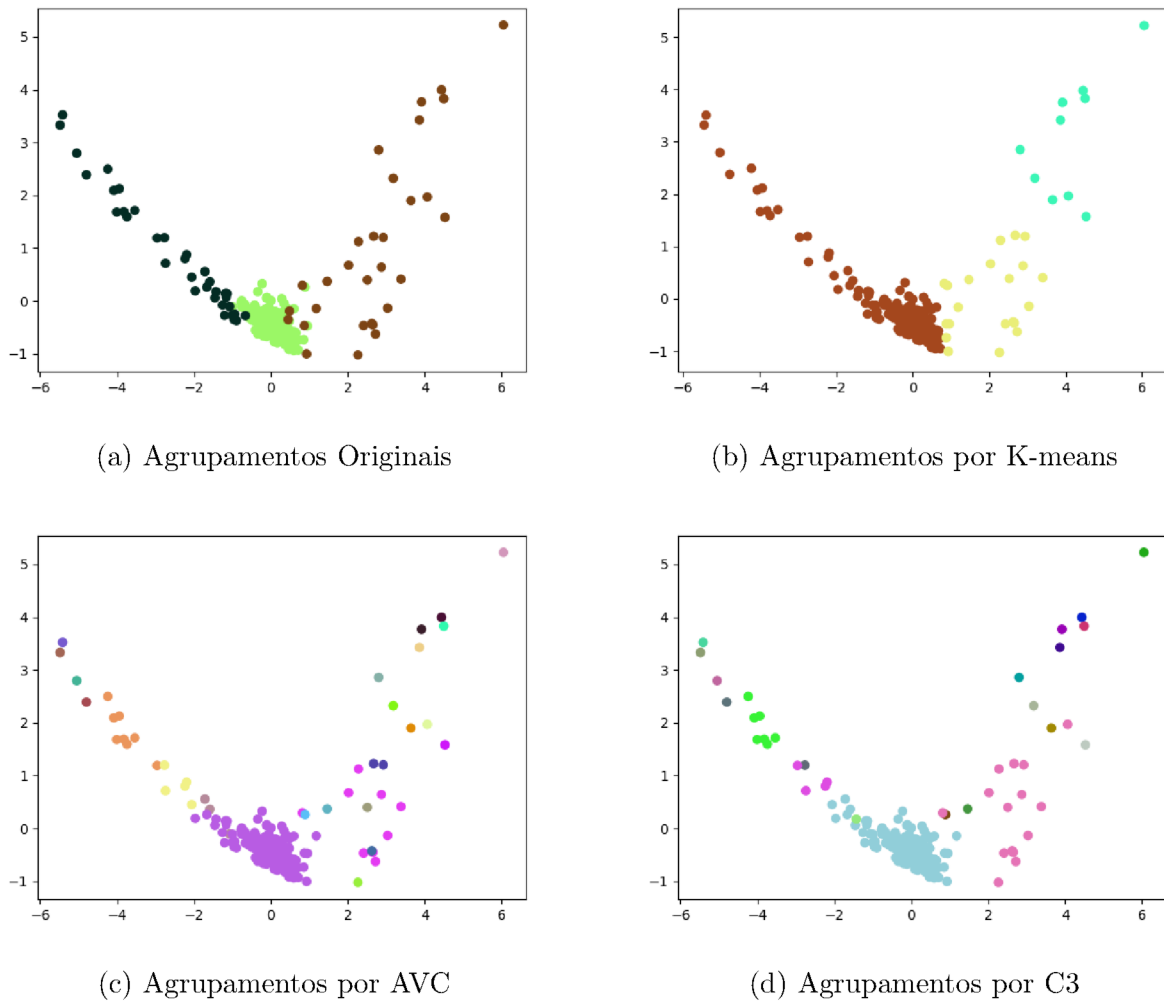


Figura 5.6

5.4 FLAME Dataset

O conjunto de dados FLAME foi criado para testes de um algoritmo de agrupamento de dados conhecido como Fuzzy clustering by Local Approximation of MEmbership (FLAME) [27]. Conta com duas coordenadas numéricas de localização e uma classe categórica. Esta subseção apresenta resultados obtidos para tarefas de agrupamento realizadas com os algoritmos K-means, AVC e C3 sobre o conjunto de dados de FLAME. Os valores apresentados para as métricas de avaliação são referentes às médias de 10 execuções consecutivas. Os hiperparâmetros utilizados neste conjunto de testes são relatados na tabela 7.

Na avaliação pelo índice de Rand ajustado, os modelos descobertos pelo algoritmo AVC apresentaram desempenho médio de 0,55, um ganho de 25% sobre o modelo descoberto pelo algoritmo K-means, com desempenho médio de 0,44. Os modelos descobertos pelo algoritmo C3 apresentaram Rand ajustado de 0,72 em média, o que representa ganhos de 63,6% e 30,9% sobre os modelos apresentados pelos algoritmos K-means e AVC, respectivamente.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apre-

Tabela 7: Parâmetros utilizados nos testes sobre o conjunto de dados FLAME.

Parâmetro	K-means	AVC	C3
k	2	-	-
Γ	-	0,7	0,7
λ	-	0,7	0,7
ϕ	-	8	8
ϵ	-	-	0,12
ϖ	-	-	34

sentou estatística- F de 213,6909 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 16,9839 e 29,1010, respectivamente.

Sob avaliação pelo índice de informação mútua, os modelos descobertos pelo algoritmo AVC apresentaram desempenho médio de 0,64, um ganho de 52,4% sobre os modelos descobertos pelo algoritmo K-means, com desempenho de 0,42 em média. Os modelos descobertos pelo algoritmo C3 apresentaram informação mútua de 0,73, em média, o que representa ganhos de 73,8% e 14,1% sobre os modelos apresentados pelos algoritmos K-means e AVC, respectivamente.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 576,7721 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 13,7828 e 46,7391, respectivamente.

A avaliação de homogeneidade de agrupamentos apresentou valor de 0,94, em média, para os modelos descobertos pelo algoritmo AVC, um ganho de 113,6% sobre os modelos descobertos pelo algoritmo K-means, com desempenho de 0,44. O modelo descoberto pelo algoritmo C3 apresentou homogeneidade de 0,93, em média, o que representa ganho de 111,4% sobre os modelos descobertos pelo algoritmo K-means, mas perda de desempenho de 1,1% em comparação com os modelos descobertos pelo algoritmo AVC.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 930,9694 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que a diferença entre a média de desempenho do algoritmos C3 e a média de desempenho do algoritmo K-means é estatisticamente significativa. A comparação da média de desempenho do algoritmo C3 com as médias de desempenho dos algoritmos AVC e K-means apresentaram estatísticas- Q de 0,6935 e 52,4978, respectivamente.

A avaliação de completude de agrupamentos apresentou valor de 0,49, em média, para os modelos descobertos pelo algoritmo AVC, o que representou ganho de desempenho de 19,5%

em comparação com os modelos descobertos pelo algoritmo K-means, com valor de 0,41. Os modelos descobertos pelo algoritmo C3 apresentaram completude de 0,61, o que representa ganhos de desempenho de 48,8% e 24,5% sobre os modelos descobertos pelos algoritmos K-means e AVC, respectivamente.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 191,7447 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 16,3044 e 27,5393, respectivamente.

Na avaliação pela métrica-V, os modelos descobertos pelo algoritmo AVC apresentaram desempenho médio de 0,64, um ganho de 52,4% sobre o modelo descoberto pelo algoritmo K-means, com desempenho de 0,42. Os modelos descobertos pelo algoritmo C3 apresentaram métrica-V média de 0,73, o que representa ganhos de 73,8% e 14,1% sobre os modelos apresentados pelos algoritmos K-means e AVC, respectivamente.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 593,4893 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 13,7991 e 47,3675, respectivamente.

Sob avaliação pelo índice de Fowlkes-Mallows, os modelos descobertos pelo algoritmo AVC apresentaram qualidade de 0,76, o que representa ganho de 4,1% sobre o desempenho do modelo descoberto pelo algoritmo K-means, com valor de 0,73. Os modelos descobertos pelo algoritmo C3 apresentaram Fowlkes-Mallows de 0,86, o que representa ganhos de 17,8% e 13,2% sobre os modelos apresentados pelos algoritmos K-means e AVC, respectivamente.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 131,3674 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 17,0655 e 21,7871, respectivamente.

O índice de Silhouette para os modelos descobertos pelo algoritmo K-means foi de 0,37, o que representa ganhos de 15,6% e 68,2% sobre os modelos descobertos pelos algoritmos AVC e C3, respectivamente. Os modelos descobertos pelo algoritmo C3 apresentaram Silhouette médio de 0,22, o que representou perda de desempenho de 31,3% sobre os modelos descobertos pelo algoritmo AVC, que obteve 0,32 para a mesma métrica.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 11,5835 e valor- p de 0,0002, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de

Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 4,6199 e 6,6393, respectivamente.

Sob avaliação pelo índice de Calinski-Harabasz, os modelos descobertos pelo algoritmo K-means apresentaram valor médio de 146,46, o que representa ganhos de desempenho de 16,72% e 89,01% sobre os desempenhos dos modelos descobertos pelos algoritmos AVC e C3, respectivamente. O algoritmo C3 apresentou modelos com Calinski-Harabasz de 77,49, enquanto o algoritmo AVC apresentou modelos de 125,48 para a mesma métrica em média, o que representa ganho de desempenho de 61,93% em favor do algoritmo AVC.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 35,3350 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 8,0685 e 11,5960, respectivamente.

O índice de Davies-Bouldin para os modelos descobertos pelo algoritmo K-means foi de 1,15 em média, o que representa perda de 74,2% sobre o modelo descoberto pelo algoritmo AVC, com valor de 0,66. Os modelos descobertos pelo algoritmo C3 apresentaram desempenho médio de 0,83 pelo índice de Davies-Bouldin, o que representa ganho de desempenho de 27,8% sobre os modelos descobertos pelo algoritmo K-means, mas perda de desempenho de 25,8% na comparação de desempenho com os modelos descobertos pelo algoritmo AVC.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 71,5866 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 5,8204 e 10,8503, respectivamente.

Os testes de Tukey para diferença honestamente significativa entre as médias de desempenho dos algoritmos C3, AVC e K-means corroboraram com a superioridade do algoritmo C3 sobre o AVC sob as métricas de qualidade de agrupamentos internas Rand ajustado, informação mútua ajustada, completude, métrica-V e Fowlkes-Mallows. A superioridade da abordagem proposta sobre o algoritmo K-means foi corroborada pelos testes de Tukey para as médias de desempenho sob as métricas internas Rand ajustado, informação mútua ajustada, homogeneidade, completude, métrica-V e Fowlkes-Mallows. O teste de Tukey para comparação das médias de desempenho sob a métrica externa de Davies-Bouldin corroborou com a superioridade de desempenho do algoritmo C3 sobre o algoritmo K-means.

A figura 5.7 apresenta uma comparação visual entre os resultados obtidos pelos modelos obtidos em média pelos 3 algoritmos utilizados nas métricas utilizadas na verificação de desempenho apresentada neste capítulo. Na figura, são destacadas as médias de resultados obtidos por algoritmo e métrica. A tabela 8 apresenta os dados utilizados na formatação dos gráficos

presentes na figura 5.7 em formato tabular.

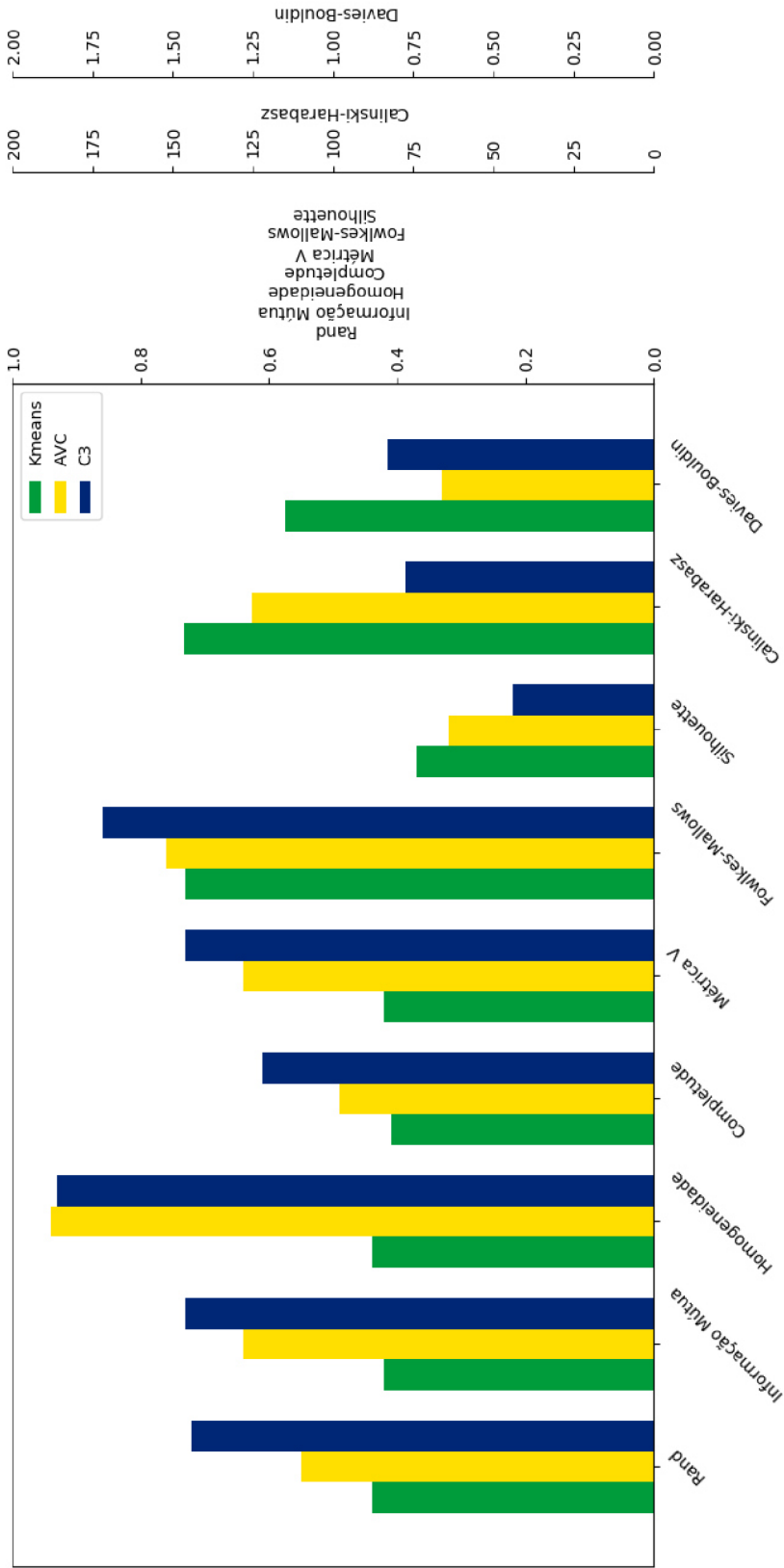
A figura 5.8 exhibe a conformação original dos agrupamentos (a), a conformação dos agrupamentos realizados pelo algoritmo K-means (b), a conformação dos agrupamentos realizados pelo algoritmo AVC (c) e a conformação dos agrupamentos realizados pelo algoritmo C3.

Na conformação original dos agrupamentos do conjunto de dados, foram calculados os graus de definição 0,32, 102,28 e 1,20, respectivamente, pelos índices internos para avaliação de agrupamentos 2 Silhouette, Calinski-Harabasz e Davies-Bouldin. Estas informações também são relacionadas à boa ou má definição das fronteiras dos agrupamentos originais do conjunto de dados analisado.

Tabela 8: Indicadores de qualidade de agrupamentos de dados (FLAME).

Algoritmo	Rand	Info Mútua	Homogeneidade	Compleitude	V	Fowlkes-Mallows	Silhouette	Calinski-Harabasz	Davies-Bouldin
K-means	0,44	0,42	0,44	0,41	0,42	0,73	0,37	146,46	1,15
AVC	0,55	0,64	0,94	0,49	0,64	0,76	0,32	125,48	0,66
C3	0,72	0,73	0,93	0,61	0,73	0,86	0,22	77,49	0,83

Figura 5.7: Comparativo visual de desempenho dos algoritmos K-means, AVC e C3 (FLAME).



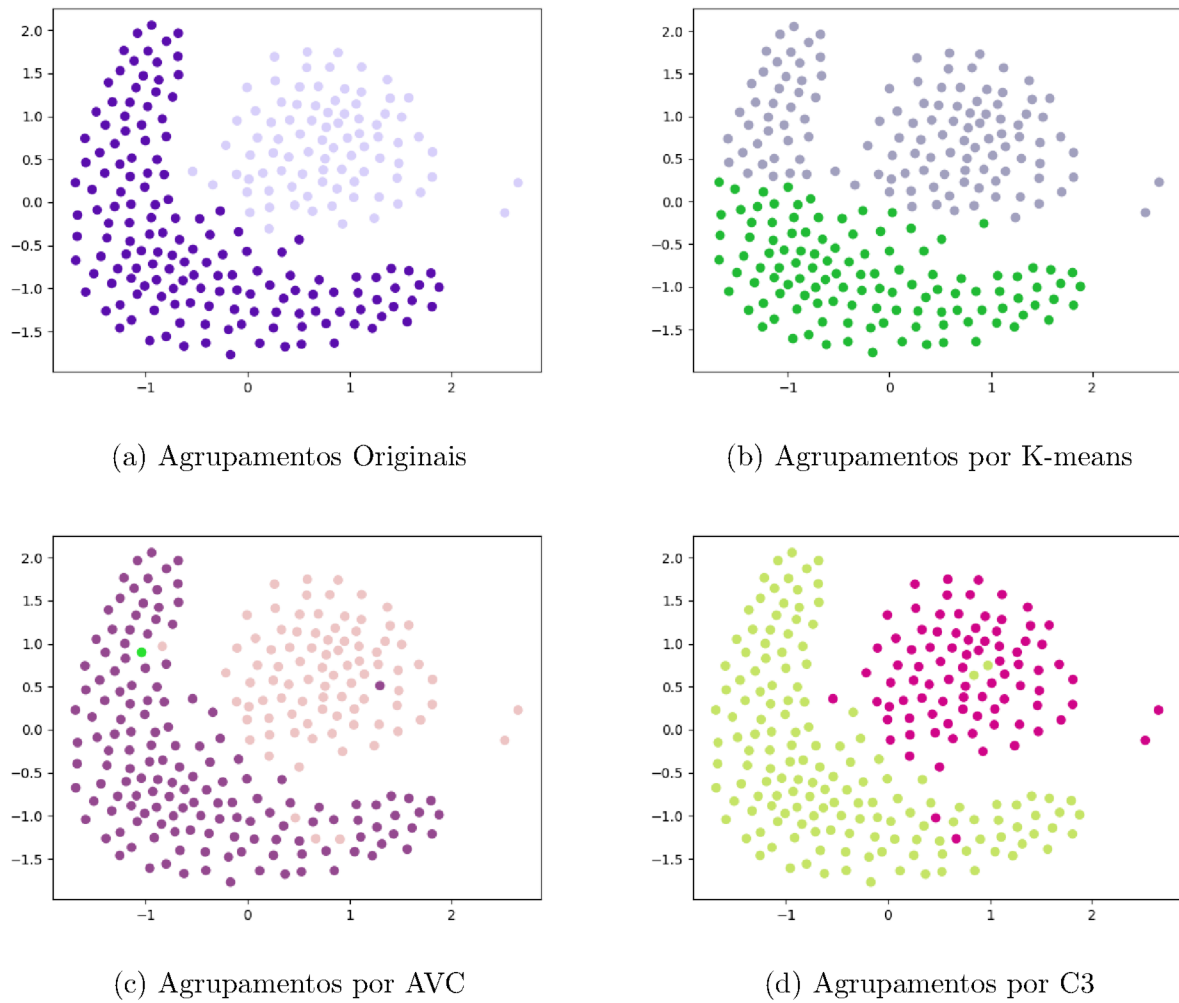


Figura 5.8

5.5 *Escherichia Coli* Dataset

O conjunto de dados de *Escherichia Coli* [49] mostra dados de localização de proteínas cuja presença no organismo humano indica infecção pela bactéria *Escherichia Coli*, comumente encontrada no estômago de animais de sangue-quente. Os agrupamentos correspondem às localizações possíveis: citoplasma, periplasma, membrana interna sem sequenciamento, membrana interna com sequenciamento indefinido, membrana externa, lipoproteína de membrana externa, membrana interna de lipoproteína e membrana interna com sequenciamento indefinido. Com exceção dos rótulos categóricos dos agrupamentos, os demais dados são numéricos.

Esta subseção apresenta resultados obtidos para tarefas de agrupamento realizadas com os algoritmos K-means, AVC e C3 sobre o conjunto de dados de localizações de proteínas em células eucariontes. Os valores apresentados para as métricas de avaliação são referentes às médias de 10 execuções consecutivas. Os hiperparâmetros utilizados neste conjunto de testes são relatados na tabela 9.

Na avaliação pelo índice de Rand ajustado, os modelos descobertos pelo algoritmo AVC

Tabela 9: Parâmetros utilizados nos testes sobre o conjunto de dados de Ecoli.

Parâmetro	K-means	AVC	C3
k	8	-	-
Γ	-	0,3	0,3
λ	-	0,3	0,3
ϕ	-	8	8
ϵ	-	-	0,12
ϖ	-	-	27

apresentaram desempenho médio de 0,53, um ganho de 47,2% sobre os modelos descobertos pelo algoritmo K-means, com desempenho médio de 0,36. Os modelos descobertos pelo algoritmo C3 apresentaram Rand ajustado de 0,59, o que representa ganhos de 63,9% e 11,3% sobre os modelos apresentados pelos algoritmos K-means e AVC, respectivamente.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 162,8379 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 5,6773 e 24,3872, respectivamente.

Sob avaliação pelo índice de informação mútua, os modelos descobertos pelo algoritmo AVC apresentaram desempenho médio de 0,48, uma perda de 7,7% sobre o modelo descoberto pelo algoritmo K-means, com desempenho de 0,52. Os modelos descobertos pelo algoritmo C3 apresentaram informação mútua de 0,50, o que representa ganho de 4,2% em comparação com os modelos descobertos pelo algoritmo AVC, mas perda de 3,9% em comparação com os modelos descobertos pelo algoritmo K-means.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 25,0602 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmos C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 6,1994 e 3,7088, respectivamente.

A avaliação de homogeneidade de agrupamentos apresentou valor médio de 0,61 para os modelos descobertos pelo algoritmo K-means, uma perda de 10,3% sobre os modelos descobertos pelo algoritmo AVC, com desempenho de 0,68. O modelo descoberto pelo algoritmo C3 apresentou homogeneidade de 0,67, o que representa ganho de 9,8% de desempenho sobre os modelos descobertos pelo algoritmo K-means, mas perda de 1,5% de desempenho sobre os modelos descobertos pelo algoritmo AVC.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 65,6157 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de

Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que a diferença entre a média de desempenho do algoritmo C3 e a média de desempenho dos algoritmo K-means é estatisticamente significativa. As estatísticas- Q para as comparações de médias de desempenho do algoritmo C3 e os algoritmos AVC e K-means foram 0,8617 e 13,5795, respectivamente.

A avaliação de completude de agrupamentos apresentou valor médio de 0,46 para os modelos descobertos pelo algoritmo AVC, o que representou perda de desempenho de 4,2% em comparação com os modelos descobertos pelo algoritmo K-means, com valor de 0,48. Os modelos descobertos pelo algoritmo C3 apresentaram completude semelhante aos modelos descobertos pelo algoritmo K-means.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 15,1957 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que a diferença entre a média de desempenho do algoritmo C3 e a média de desempenho do algoritmo AVC é estatisticamente significativa. As comparações de médias de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means apresentaram estatísticas- Q de 7,5136 e 1,9549, respectivamente.

Na avaliação pela métrica-V, os modelos descobertos pelo algoritmo AVC apresentaram desempenho médio de 0,54, semelhante ao desempenho médio dos modelos descobertos pelo algoritmo K-means. Os modelos descobertos pelo algoritmo C3 apresentaram métrica-V média de 0,56, o que representa ganhos de 3,7% sobre os modelos descobertos pelos algoritmos K-means e AVC.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 15,6586 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmo C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 5,3825 e 7,7159, respectivamente.

Sob avaliação pelo índice de Fowlkes-Mallows, os modelos descobertos pelo algoritmo AVC apresentaram qualidade média de 0,65, o que representa ganho de 30,0% sobre o desempenho médio dos modelos descobertos pelo algoritmo K-means, com valor de 0,50. Os modelos descobertos pelo algoritmo C3 apresentaram Fowlkes-Mallows médio de 0,69, o que representa ganhos de 38,0% e 6,2% sobre os modelos apresentados pelos algoritmos K-means e AVC, respectivamente.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 161,2659 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmo C3 e as médias de desempenho dos algoritmos AVC e K-means são estatisticamente significantes, com estatísticas- Q de 5,5516 e 24,2393, respectivamente.

O índice de Silhouette médio para os modelos descobertos pelo algoritmo AVC foi de -0,18,

o que representa perda de 151,4% sobre os modelos descobertos pelo algoritmo K-means, que obtiveram valor médio de 0,35 para a mesma métrica. Os modelos descobertos pelo algoritmo C3 apresentaram Silhouette de -0,17, o que representou ganho de desempenho de 5,6% sobre os modelos descobertos pelo algoritmo AVC, mas perda de desempenho de 148,6% em comparação com os modelos descobertos pelo algoritmo K-means.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 317,7211 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmo C3 e a média de desempenho do algoritmo K-means é estatisticamente significativa. As comparações de médias de desempenho do algoritmo C3 e os algoritmos AVC e K-means apresentaram estatísticas- Q de 0,2033 e 30,7712, respectivamente.

Sob avaliação pelo índice de Calinski-Harabasz, os modelos descobertos pelo algoritmo K-means apresentaram valor médio de 276,15, o que representa ganho de desempenho de 254,8% sobre os modelos descobertos pelo algoritmo AVC, que apresentou o valor médio 77,84 para a mesma métrica. O algoritmo C3 apresentou modelos com Calinski-Harabasz de 74,83, em média, o que representa perdas de desempenho de 72,9% e 3,9% em comparação com os modelos descobertos pelos algoritmos K-means e AVC, respectivamente.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 2.142,9864 e valor- $p < 0,0001$, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmo C3 e a média de desempenho do algoritmo K-means é estatisticamente significativa. As comparações de médias de desempenho do algoritmo C3 e os algoritmos AVC e K-means apresentaram estatísticas- Q de 1,2072 e 80,7776, respectivamente.

O índice de Davies-Bouldin para os modelos descobertos pelo algoritmo AVC teve valor médio de 0,80, o que representou ganho de desempenho de 16,7% sobre os modelos descobertos pelo algoritmo K-means, cujo valor para a mesma métrica foi 0,96. O algoritmo C3 apresentou modelos com Davies-Bouldin de 0,84, o que representou ganho de desempenho de 12,5% na comparação com os modelos descobertos pelo algoritmo K-means, mas perda de 5,0% de desempenho na comparação com os modelos descobertos pelo algoritmo AVC.

A ANOVA simples dos dados utilizados para os cálculos de médias de desempenho apresentou estatística- F de 4,224 e valor- p de 0,0253, sugerindo que uma ou mais médias de desempenho são significativamente distintas das demais. A aplicação subsequente do teste de Tukey com estatística- $Q_{\text{crítica}}$ 3,5058 revelou que as diferenças entre a média de desempenho do algoritmo C3 e as médias de desempenhos dos algoritmos AVC e K-means são estatisticamente insignificantes. As comparações de médias de desempenho do algoritmo C3 e os algoritmos AVC e K-means apresentaram estatísticas- Q de 0,9034 e 3,0212, respectivamente.

Os testes de Tukey para diferença honestamente significativa entre as médias de desempenho dos algoritmos C3, AVC e K-means corroboraram com a superioridade do algoritmo C3 sobre

o AVC sob as métricas de qualidade de agrupamentos internas Rand ajustado, informação mútua ajustada, completude, métrica-V e Fowlkes-Mallows. A superioridade da abordagem proposta sobre o algoritmo K-means foi corroborada por testes de Tukey sobre as médias de desempenho dos dois algoritmos sob as métricas internas Rand ajustado, homogeneidade, métrica-V e Fowlkes-Mallows.

A figura 5.9 apresenta uma comparação visual entre os resultados obtidos pelos modelos obtidos em média pelos 3 algoritmos utilizados nas métricas utilizadas na verificação de desempenho apresentada neste capítulo. Na figura, são destacadas as médias de resultados obtidos por algoritmo e métrica. A tabela 10 apresenta os dados utilizados na formatação dos gráficos presentes na figura 5.9 em formato tabular.

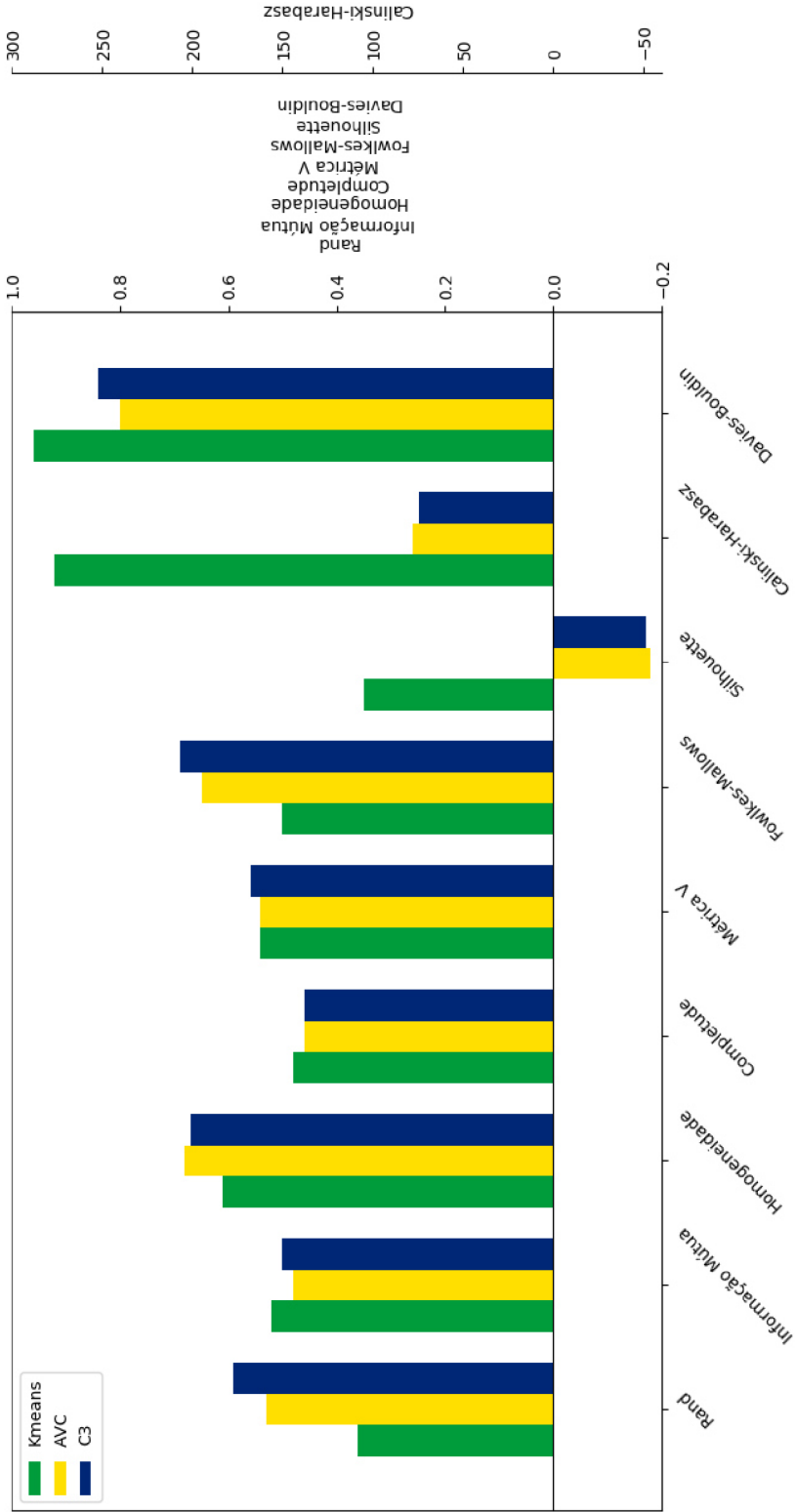
A figura 5.10 exibe a conformação original dos agrupamentos (a), a conformação dos agrupamentos realizados pelo algoritmo K-means (b), a conformação dos agrupamentos realizados pelo algoritmo AVC (c) e a conformação dos agrupamentos realizados pelo algoritmo C3.

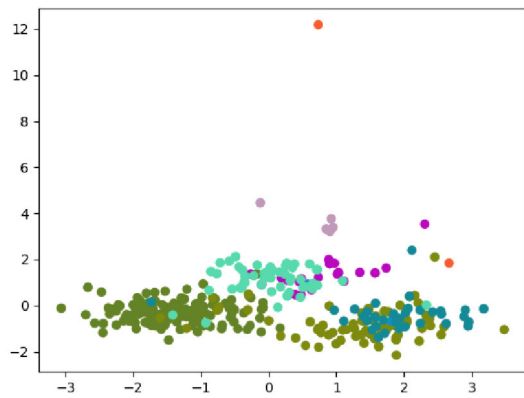
Na conformação original dos agrupamentos do conjunto de dados, foram calculados os graus de definição 0,23, 120,98 e 1,85, respectivamente, pelos índices internos para avaliação de agrupamentos 2 Silhouette, Calinski-Harabasz e Davies-Bouldin. Estas informações também são relacionadas à boa ou má definição das fronteiras dos agrupamentos originais do conjunto de dados analisado.

Tabela 10: Indicadores de qualidade de agrupamentos de dados (*Escherichia Coli*).

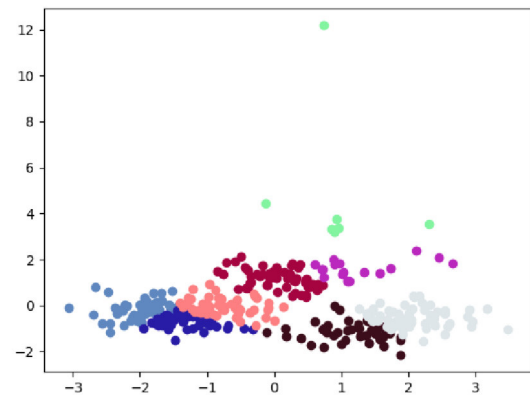
Algoritmo	Rand	Info Mútua	Homogeneidade	Compleitude	V	Fowlkes-Mallows	Silhouette	Calinski-Harabasz	Davies-Bouldin
K-means	0,36	0,52	0,61	0,48	0,54	0,50	0,35	276,15	0,96
AVC	0,53	0,48	0,68	0,46	0,54	0,65	-0,18	77,84	0,80
C3	0,59	0,50	0,67	0,46	0,56	0,69	-0,17	74,83	0,84

Figura 5.9: Comparativo visual de desempenho dos algoritmos K-means, AVC e C3 (*Escherichia Coli*).

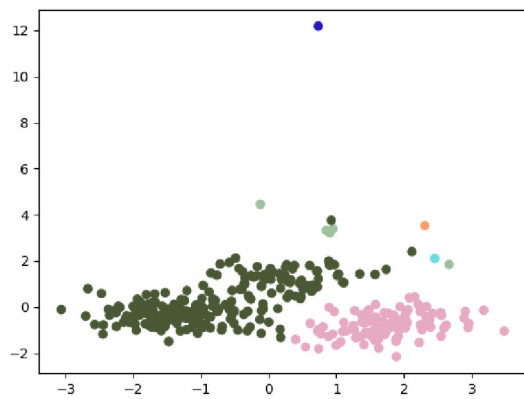




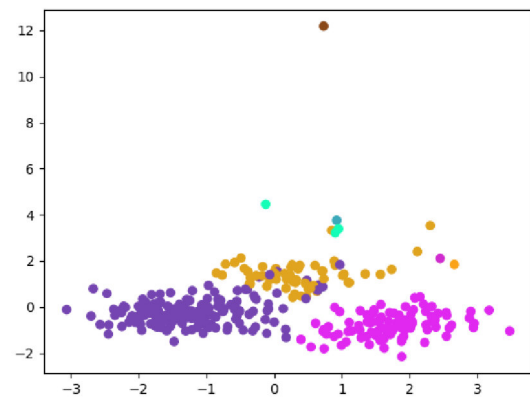
(a) Agrupamentos Originais



(b) Agrupamentos por K-means



(c) Agrupamentos por AVC



(d) Agrupamentos por C3

Figura 5.10

6 DISCUSSÃO

Este capítulo apresenta a análise dos resultados obtidos na comparação de desempenho dos algoritmos C3, AVC e K-means, conforme os experimentos descritos no capítulo 5. No total, foram realizadas 50 execuções de algoritmos, sendo 10 para cada algoritmo sobre cada conjunto de dados analisado.

O algoritmo C3 apresentou desempenho superior aos algoritmos K-means e AVC, de acordo com o índice de Rand ajustado, em 9 dos 10 cenários estudados. Em média, a diferença entre C3 e K-means foi de 27% em favor da abordagem proposta. A diferença na comparação com o algoritmo AVC foi, em média, 16%, também em favor do algoritmo C3. Em consideração do índice de informação mútua, o algoritmo C3 mostrou superioridade de desempenho em 5 dos 10 cenários estudados. Em média, a diferença entre C3 e K-means foi de 2% em favor do primeiro. A diferença na comparação da abordagem proposta com o algoritmo AVC foi, em média, 6% em favor do C3.

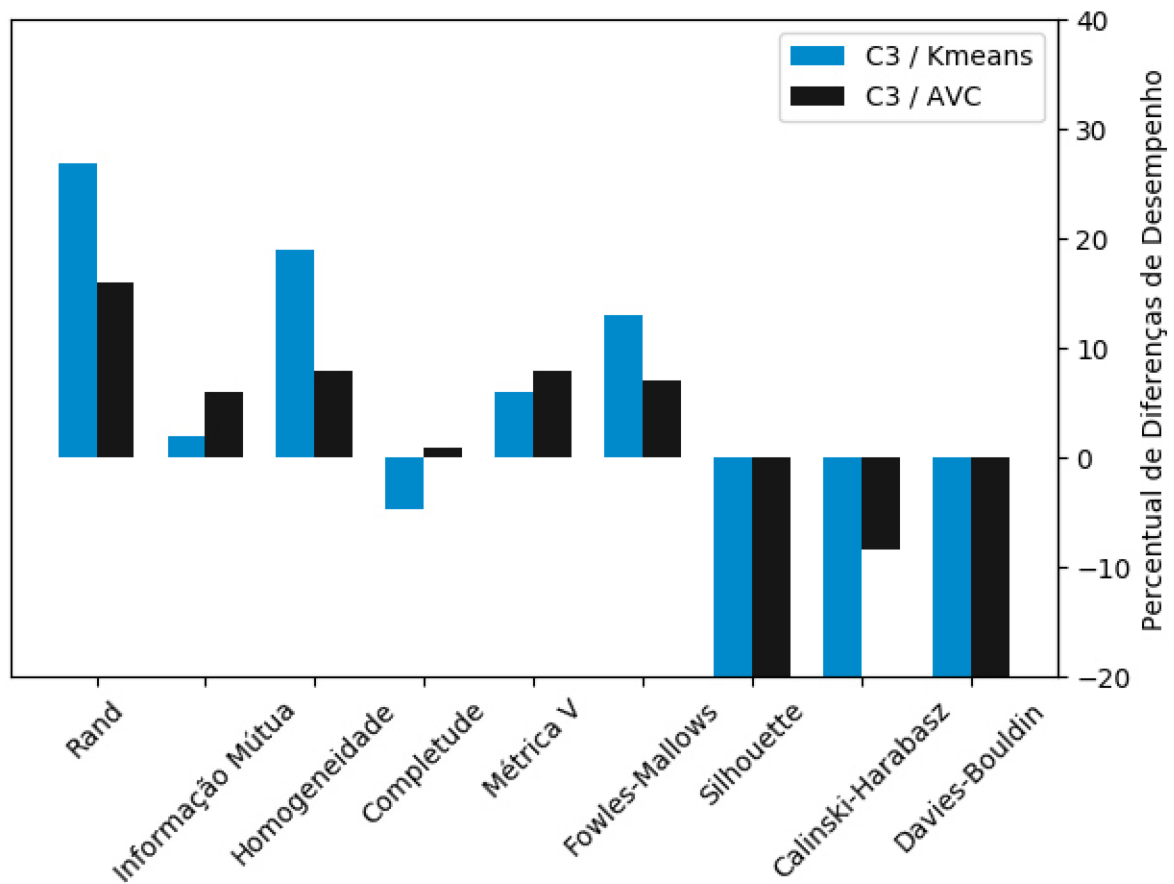
Em termos de homogeneidade, o algoritmo C3 mostrou superioridade de desempenho em 6 dos 10 cenários estudados. Em média, a diferença entre C3 e K-means foi de 19% em favor do primeiro, enquanto a diferença entre C3 e AVC foi de 8%, também em favor do C3. Sob o critério de completude, o algoritmo C3 mostrou superioridade em 5 dos 10 cenários estudados. Em média, a diferença entre C3 e K-means foi 4,6% em favor deste, enquanto a diferença entre C3 e AVC foi de 1% em favor do primeiro.

A métrica V corroborou com a superioridade do algoritmo C3 sobre os algoritmos K-means e AVC em 7 dos 10 cenários estudados. As diferenças entre os desempenhos dos algoritmos C3, K-means e AVC foi, em média, de 6% e 8%, respectivamente. A análise pelo índice de Fowlkes-Mallows mostrou superioridade de desempenho do algoritmo C3 sobre os algoritmos K-means e AVC em 9 dos 10 cenários avaliados, com diferenças de 13% e 7%, respectivamente.

O índice de Silhouette mostrou superioridade de desempenho do algoritmo C3 sobre os algoritmos K-means e AVC em 2 dos 10 cenários estudados. Contudo, as médias de desempenho mostraram inferioridade do algoritmo C3 de 26,2% na comparação com o algoritmo K-means e 59,9% na comparação com o algoritmo AVC. O índice de Calinski-Harabasz mostrou superioridade de desempenho do algoritmo C3 em 2 dos 10 cenários estudados. As médias de diferenças de desempenho entre o algoritmo C3 e os algoritmos AVC e K-means foram 8,4% e 57,9%, respectivamente, em favor destes. O índice de Davies-Bouldin mostrou superioridade de desempenho do algoritmo C3 sobre os algoritmos K-means e AVC em 5 dos 10 cenários estudados. As médias de desempenho mostraram perda de desempenho do C3 em comparação com os algoritmos AVC e K-means de 75,6% e 57,3%, respectivamente.

A figura 6.1 apresenta um comparativo gráfico entre os ganhos e perdas de desempenho do algoritmo C3 em relação aos algoritmos K-means e AVC.

Figura 6.1: Comparativo de Diferenças de Desempenho Médio do algoritmo C3 sobre os algoritmos K-means e AVC.



Os resultados obtidos mostraram que, no geral, sob avaliação por métricas internas de qualidade de agrupamentos, a abordagem proposta, o algoritmo C3, teve superioridade de desempenho de 10,4% sobre o algoritmo K-means, um dos algoritmos mais populares e aplicados em tarefas de agrupamento de dados; e de 7,7% sobre o algoritmo AVC, que pode ser compreendido como uma variante do algoritmo C3 que não é capaz de avaliar aspectos cognitivos, mas apenas de densidade populacional. Por outro lado, quando da avaliação de agrupamentos por métricas externas, na média dos 3 índices utilizados, o algoritmo C3 mostrou inferioridade de 58% sobre o algoritmo K-means, embora tenha apresentado superioridade de 36,7%, em média, na comparação com o algoritmo AVC.

7 CONCLUSÃO

O presente trabalho abordou o problema do agrupamento de dados sob uma metodologia original que, fundamentalmente, integra modelos de consistência cognitiva aos processos de tomada de decisão que impactam precisamente no pertencimento dos pontos de dados analisados aos agrupamentos disponíveis no sistema.

O objetivo geral do trabalho, da investigação do impacto da aplicação de modelos de processos cognitivos humanos em tarefas de agrupamentos de dados e como isto influencia na qualidade dos modelos descobertos foi cumprido a partir da proposta de um algoritmo de agrupamento de dados original. O algoritmo Cognitive Consistency Clustering (C3) deu realização ao primeiro objetivo específico do trabalho. Ele é capaz de representar conjuntos de dados na forma de um sistema multiagente básico, onde os elementos são dotados de cognições, por vezes dissonantes, que pautam suas tomadas de decisão pela direção dos próximos movimentos dos indivíduos do sistema. Tais movimentações acontecem até que o sistema encontre estabilidade na conformação dos agrupamentos de dados.

Foram realizados experimentos de aprendizado não supervisionado com a abordagem proposta e outras relacionadas para verificação de desempenho e validação da hipótese, cujos resultados revelaram que métricas de avaliação internas propõem melhor desempenho na descoberta de agrupamentos do que as métricas externas propõem.

Entende-se que isto é consequência do método de tomada de decisão para movimentação dos agentes do sistema, que, dotados da capacidade de avaliação da consistência cognitiva das opções disponíveis, podem descobrir relações interagentes que seriam preteridas ou simplesmente não descobertas por abordagens cujas tomadas de decisão são baseadas unicamente em densidade populacional. Isto evidencia o cumprimento do objetivo específico da verificação de superioridade de desempenho de acurácia e definição de agrupamentos, na comparação com algoritmos relacionados.

A hipótese de que a aplicação de teorias do balanceamento psicológico e dissonância cognitiva no contexto apresentado tem o potencial de resultar em modelos de melhor desempenho, em comparação com modelos gerados por métodos tradicionais, foi, portanto, avaliada como verdadeira.

Entretanto, não foram observadas melhorias de desempenho pela abordagem proposta quando os agrupamentos a serem descobertos apresentam fronteiras bem definidas entre si, o que comumente é caracterizado em conjuntos de dados sintéticos. A principal contribuição deste trabalho ocorre em cenários em que os agrupamentos a serem descobertos apresentam fronteiras mal definidas, ou que não sejam caracterizadas linearmente, o que é comumente verificado em conjuntos de dados oriundos de situações do mundo real.

Esta avaliação, por fim, direciona a continuidade do trabalho, que envolve a prospecção de parâmetros que possam eventualmente ser utilizados no melhoramento de desempenho sob óptica de métricas externas de qualidade de agrupamentos. Outro aspecto que norteia os

trabalhos futuros é a investigação mecanismos lógicos que permitam a redução de complexidade computacional da proposta, visando a escalabilidade do algoritmo desenvolvido sem necessidade da aplicação de técnicas redução de dimensionalidade de dados.

8 REFERÊNCIAS

- [1] Hervé Abdi and Lynne J Williams. Tukey’s honestly significant difference (hsd) test. *Encyclopedia of research design*, 3(1):1–5, 2010.
- [2] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 94–105, 1998.
- [3] Ziv Bar-Joseph, David K Gifford, and Tommi S Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17(suppl_1):S22–S29, 2001.
- [4] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.
- [5] Gerald Berman and Kenneth D Fryer. *Introduction to combinatorics*. Elsevier, 2014.
- [6] James C Bezdek. Objective function clustering. In *Pattern recognition with fuzzy objective function algorithms*, pages 43–93. Springer, 1981.
- [7] S. Bhatti, C. Beck, and A. Nedić. Data clustering and graph partitioning via simulated mixing. *IEEE Transactions on Network Science and Engineering*, 6(3):253–266, 2019.
- [8] Jack W Brehm. Postdecision changes in the desirability of alternatives. *The Journal of Abnormal and Social Psychology*, 52(3):384, 1956.
- [9] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.
- [10] Dorwin Cartwright and Frank Harary. Structural balance: a generalization of heider’s theory. *Psychological review*, 63(5):277, 1956.
- [11] M. Chen, X. Wen, Z. Yang, M. Li, and M. Zhang. Multisim: A novel similar-to-multiple-point clustering algorithm. *IEEE Access*, 6:78225–78237, 2018.
- [12] Z. Chen, D. Chang, and Y. Zhao. An automatic clustering algorithm based on region segmentation. *IEEE Access*, 6:74247–74259, 2018.
- [13] Sandra CM Cohen and Leandro Nunes de Castro. Data clustering with particle swarms. In *2006 IEEE International Conference on Evolutionary Computation*, pages 1792–1798. IEEE, 2006.
- [14] D Coomans, I Broeckaert, M Jonckheer, and DL Massart. Comparison of multivariate discrimination techniques for clinical data—application to the thyroid functional state. *Methods of Information in Medicine*, 22(02):93–101, 1983.

- [15] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.
- [16] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.
- [17] Chris Ding. *Dimension Reduction Techniques for Clustering*, pages 846–846. Springer US, Boston, MA, 2009.
- [18] JC Dunn. A graph theoretic analysis of pattern classification via tamura’s fuzzy relation. *IEEE Transactions on Systems, Man, and Cybernetics*, (3):310–313, 1974.
- [19] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.
- [20] Leon Festinger. *A theory of cognitive dissonance*, volume 2. Stanford university press, 1957.
- [21] Leon Festinger. Conflict, decision, and dissonance. 1964.
- [22] Claude Flament. *Applications of graph theory to group structure*. Prentice-Hall, 1963.
- [23] M Forina. An extendible package for data exploration, classification and correlation (1998). *Institute of Pharmaceutical and Food Analysis and Technologies*.
- [24] RS Forsyth. Pc/beagle user’s guide. *BUPA Medical Research Ltd*, 1990.
- [25] Edward B Fowlkes and Colin L Mallows. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569, 1983.
- [26] Heider Fritz et al. The psychology of interpersonal relations, 1958.
- [27] Limin Fu and Enzo Medico. Flame, a novel fuzzy clustering method for the analysis of dna microarray data. *BMC bioinformatics*, 8(1):1–15, 2007.
- [28] Guojun Gan, Chaoqun Ma, and Jianhong Wu. *Data clustering: theory, algorithms, and applications*. SIAM, 2007.
- [29] Lutiele M. Godois, Diana F. Adamatti, and Leonardo R. Emmendorfer. A multi-agent-based algorithm for data clustering. *Progress in Artificial Intelligence*, 9(3):218, 2020.
- [30] P. Grabusts and A. Borisov. Using grid-clustering methods in data classification. In *Proceedings. International Conference on Parallel Computing in Electrical Engineering*, pages 425–426, 2002.

- [31] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Cure: an efficient clustering algorithm for large databases. *ACM Sigmod record*, 27(2):73–84, 1998.
- [32] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Rock: A robust clustering algorithm for categorical attributes. *Information systems*, 25(5):345–366, 2000.
- [33] Xiaofei He and Partha Niyogi. Locality preserving projections. *Advances in neural information processing systems*, 16, 2003.
- [34] David C Howell. *Statistical methods for psychology*. Cengage Learning, 2012.
- [35] L Hubert and P Arabie. Comparing clusterings. *Journal of Classification*, 2:193–218, 1985.
- [36] Anil K Jain and Richard C Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.
- [37] George Karypis, Eui-Hong Han, and Vipin Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75, 1999.
- [38] Leonard Kaufmann. Clustering by means of medoids. In *Proc. Statistical Data Analysis Based on the L1 Norm Conference, Neuchatel, 1987*, pages 405–416, 1987.
- [39] Eamonn Keogh, Kaushik Chakrabarti, Michael Pazzani, and Sharad Mehrotra. Locally adaptive dimensionality reduction for indexing large time series databases. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, pages 151–162, 2001.
- [40] Hae-Young Kim. Analysis of variance (anova) comparing means of more than two groups. *Restorative dentistry & endodontics*, 39(1):74–77, 2014.
- [41] S. Li, L. Li, J. Yan, and H. He. Sde: A novel clustering framework based on sparsity-density entropy. *IEEE Transactions on Knowledge and Data Engineering*, 30(8):1575–1587, 2018.
- [42] S. Liang and D. Han. A clustering method based on neighborhood chain. In *2018 Chinese Automation Congress (CAC)*, pages 2831–2835, 2018.
- [43] H. Liu, J. Cheng, and F. Wang. Sequential subspace clustering via temporal smoothness. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 245–250, 2017.
- [44] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.
- [45] Thales Vaz Maciel and Leonardo Ramos Emmendorfer. Cognitive consistency models applied to data clustering. In Leszek Rutkowski, Rafal Scherer, Marcin Korytkowski, Witold Pedrycz, Ryszard Tadeusiewicz, and Jacek M. Zurada, editors, *Artificial Intelligence and*

Soft Computing - 20th International Conference, ICAISC 2021, Virtual Event, June 21-23, 2021, Proceedings, Part II, volume 12855 of *Lecture Notes in Computer Science*, pages 183–191. Springer, 2021.

- [46] E. M. Mansour and A. Ahmadi. A novel clustering algorithm based on fully-informed particle swarm. In *2019 IEEE Congress on Evolutionary Computation (CEC)*, pages 713–720, 2019.
- [47] Leslie C Morey and Alan Agresti. The measurement of classification agreement: An adjustment to the rand statistic for chance agreement. *Educational and Psychological Measurement*, 44(1):33–37, 1984.
- [48] A. Mortezaezhad and E. Daneshifar. Big-data clustering with genetic algorithm. In *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)*, pages 702–706, 2019.
- [49] Kenta Nakai and Minoru Kanehisa. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, 14(4):897–911, 1992.
- [50] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.
- [51] L. Ping and R. Xiangsheng. A new clustering algorithm based on rewarding support vectors. In *2019 2nd International Conference on Information Systems and Computer Aided Education (ICISCAE)*, pages 630–634, 2019.
- [52] H. Qu, J. Wu, and Z. Li. A new clustering algorithm inspired by insect pollination. In *2019 Chinese Automation Congress (CAC)*, pages 1293–1298, 2019.
- [53] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.
- [54] P. Rastin, B. Matei, G. Cabanes, and I. E. Baghdadi. Signal-based autonomous clustering for relational data. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 4103–4110, 2017.
- [55] P. Rathore, D. Kumar, J. C. Bezdek, S. Rajasegarar, and M. Palaniswami. A rapid hybrid clustering algorithm for large volumes of high dimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 31(4):641–654, 2019.
- [56] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420, 2007.

- [57] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
- [58] Hyeong-cheol Ryu, Sungwon Jung, and Sakti Pramanik. An effective clustering method over cf+ tree using multiple range queries. *IEEE Transactions on Knowledge and Data Engineering*, 2019.
- [59] S. Sarfraz, V. Sharma, and R. Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8926–8935, 2019.
- [60] E. Schikuta. Grid-clustering: an efficient hierarchical clustering method for very large data sets. In *Proceedings of 13th International Conference on Pattern Recognition*, volume 2, pages 101–105 vol.2, 1996.
- [61] Shokri Z Selim and Mohamed A Ismail. K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on pattern analysis and machine intelligence*, (1):81–87, 1984.
- [62] H. Shao, P. Zhang, X. Chen, F. Li, and G. Du. A hybrid and parameter-free clustering algorithm for large data sets. *IEEE Access*, 7:24806–24818, 2019.
- [63] Heng Tao Shen. *Dimensionality Reduction*, pages 843–846. Springer US, Boston, MA, 2009.
- [64] Mark J Shensa et al. The discrete wavelet transform: wedding the a trous and mallat algorithms. *IEEE Transactions on signal processing*, 40(10):2464–2482, 1992.
- [65] Kristina P Sinaga and Miin-Shen Yang. Unsupervised k-means clustering algorithm. *IEEE access*, 8:80716–80727, 2020.
- [66] Z. Su and T. Denoeux. Bpec: Belief-peaks evidential clustering. *IEEE Transactions on Fuzzy Systems*, 27(1):111–123, 2019.
- [67] L. Sun, R. Liu, J. Xu, S. Zhang, and Y. Tian. An affinity propagation clustering method using hybrid kernel function with lle. *IEEE Access*, 6:68892–68909, 2018.
- [68] John W Tukey. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114, 1949.
- [69] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854, 2010.

- [70] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayоров, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.
- [71] J. Wang, L. You, W. Fan, F. Miao, and T. Yang. dfc: A data-density-aware fuzzy clustering algorithm for imbalanced biomedical datasets. In *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 563–566, 2017.
- [72] R. Wang, J. Zhou, X. Liu, S. Han, L. Wang, and Y. Chen. Transfer clustering based on gaussian mixture model. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2522–2526, 2019.
- [73] Wei Wang, Jiong Yang, Richard Muntz, et al. Sting: A statistical information grid approach to spatial data mining. In *VLDB*, volume 97, pages 186–195, 1997.
- [74] Y. Wang, J. Dong, J. Zhou, L. Wang, S. Han, T. Zhang, and C. L. P. Chen. Spectral clustering based on js-divergence for uncertain data. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1972–1975, 2017.
- [75] Shmuel Winograd. On computing the discrete fourier transform. *Mathematics of computation*, 32(141):175–199, 1978.
- [76] Kyoung-Gu Woo, Jeong-Hoon Lee, Myoung-Ho Kim, and Yoon-Joon Lee. Findit: a fast and intelligent subspace clustering algorithm using dimension voting. *Information and Software Technology*, 46(4):255–271, 2004.
- [77] C. Wu, J. Lee, T. Isokawa, J. Yao, and Y. Xia. Efficient clustering method based on density peaks with symmetric neighborhood relationship. *IEEE Access*, 7:60684–60696, 2019.
- [78] X. Yan, A. Homaifar, S. Nazmi, and M. Razeghi-Jahromi. A novel clustering algorithm based on fitness proportionate sharing. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1960–1965, 2017.
- [79] L. Yang, K. Li, W. Zhang, Z. Ke, K. Xiao, and Z. Du. An improved chaotic aco clustering algorithm. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1642–1649, 2018.

- [80] R. Zhang, L. Zhu, T. Du, S. Qu, and X. Wang. A robust density clustering algorithm based on gravity peak. In *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 544–549, 2019.
- [81] N. Zhao, L. Zhang, B. Du, Q. Zhang, J. You, and D. Tao. Robust dual clustering with adaptive manifold regularization. *IEEE Transactions on Knowledge and Data Engineering*, 29(11):2498–2509, 2017.
- [82] W. Zhao and C. Lian. A novel ensemble clustering approach with internal weighting strategy. In *2019 Chinese Automation Congress (CAC)*, pages 2521–2526, 2019.
- [83] J. Zhong, Y. Wang, H. Du, and W. Tong. A new clustering algorithm by using boundary information. In *2018 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8, 2018.
- [84] W. Zhu, W. Luo, L. Ni, and N. Lu. Swarm clustering algorithm: Let the particles fly for a while. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1242–1249, 2018.