CLEITON RODRIGUES MONTEIRO

CLASSIFICAÇÃO ESTRUTURAL DE PROTEÍNAS POR MEIO DE APRENDIZADO NÃO SUPERVISIONADO

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS − BRASIL
2019

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
M776c
2019

Monteiro, Cleiton Rodrigues, 1981-
Classificação estrutural de proteínas por meio de aprendizado não supervisionado / Cleiton Rodrigues Monteiro. – Viçosa, MG, 2019.
xiv, 79 f. : il. (algumas color.) ; 29 cm.

Inclui apêndices.
Orientador: Sabrina de Azevedo Silveira.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 43-46.

1. Bioinformática. 2. Proteínas - Estrutura. 3. Análise por agrupamento. I. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. II. Título.

CDD 22. ed. 572.60285

CLEITON RODRIGUES MONTEIRO

CLASSIFICAÇÃO ESTRUTURAL DE PROTEÍNAS POR MEIO DE APRENDIZADO NÃO SUPERVISIONADO

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

APROVADA: 25 de junho de 2019.

Jugurta Lisboa Filho

Adriana Maria Patarroyo Vargas

Sabrina de Azevedo Silveira
(Orientadora)

À família e aos amigos.

"Tudo posso naquele que me fortalece."
(Filipenses 4:13)

Agradecimentos

Agradeço primeiramente a Deus por mais uma conquista e por ter permitido que eu seguisse, mesmo nos momentos mais difíceis.

Agradeço à minha família: minha mãe pela preocupação nas frequentes viagens, meu irmão pelo incentivo de sempre, minha esposa pela paciência e dedicação, e minha filha, que mesmo ainda não entendendo o significado de tudo isso, foi minha maior motivação para continuar seguindo.

Agradeço à professora Sabrina de Azevedo Silveira pela parceria e por orientar-me sabiamente ao longo do caminho. Agradeço também ao professor Giovanni Ventorim Comarela por suas importantes contribuições.

Agradeço à Universidade Federal de Viçosa, ao Departamento de Informática e a todos os professores com os quais tive a oportunidade de aprender. Agradeço aos colegas de curso, em especial ao colega Vinício Fragoso Mendes, o qual tem uma importante participação neste trabalho.

Agradeço ao Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais - IF Sudeste MG, *Campus* Manhuaçu.

E finalmente, agradeço a todos aqueles que contribuíram direta ou indiretamente para que eu chegasse até aqui.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

MONTEIRO, Cleiton Rodrigues, M.Sc., Universidade Federal de Viçosa, junho de 2019. **Classificação estrutural de proteínas por meio de aprendizado não supervisionado**. Orientadora: Sabrina de Azevedo Silveira. Coorientador: Giovanni Ventorim Comarela.

A bioinformática estrutural se dedica ao estudo das estruturas tridimensionais de proteínas e macromoléculas. Neste trabalho, o interesse está nas estruturas de proteínas. A disponibilização de novas sequências e estruturas proteicas em bases públicas de dados tem ocorrido em um ritmo bastante acelerado, aumentando também a necessidade de métodos automáticos e eficientes para a extração e compreensão desse grande volume de dados. Segundo Gao et al. [2018], a bioinformática é uma ciência de mineração e interpretação de dados biológicos. Para eles, o fluxo contínuo e crescente desses dados, assim como a necessidade de abordar problemas biomédicos cada vez mais complexos, tem gerado oportunidades desafiadoras para pesquisadores de mineração de dados e aprendizagem de máquina. Diversas estratégias para classificação estrutural de proteínas têm sido propostas nos últimos anos, utilizando descritores baseados em sequência e estrutura. Nesta pesquisa, avaliou-se a possibilidade de classificação estrutural de proteínas utilizando métodos não supervisionados associados a características propostas com sucesso em um classificador estrutural bem estabelecido. Foram realizados experimentos utilizando 5 algoritmos de agrupamento de 4 diferentes paradigmas. A qualidade dos grupos foi avaliada por meio do Coeficiente de Silhueta e os rótulos previstos foram comparados às classes e superfamílias da base CATH, por meio do índice *Fowlkes Mallows* e da verificação

de homogeneidade e completude dos grupos. Os resultados mostram a inviabilidade de classificação no nível classe, já que os índices alcançados com *Fowlkes Mollows* não chegaram a 60%. Por outro lado, eles indicam uma capacidade considerável de classificação no nível superfamília - foi alcançado com o método *Complete-Link* um índice superior a 70% no agrupamento geral. Os resultados são ainda mais interessantes quando restringe-se o número de grupos, alcançando um índice de 78.5% para topologias com até 25 superfamílias e de 82.8% para topologias com até 5 superfamílias. Se considerados ainda, agrupamentos com índice igual ou superior a 85%, eles representam aproximadamente 40% das topologias utilizadas, sendo que deste grupo, quase metade dos agrupamentos (48.19%) obteve um índice de 100% de similaridade, ou seja, em cerca de 20% das topologias, todas as proteínas foram agrupadas corretamente.

Abstract

MONTEIRO, Cleiton Rodrigues, M.Sc., Universidade Federal de Viçosa, June, 2019. **Protein structural classification through unsupervised learning**. Advisor: Sabrina de Azevedo Silveira. Co-Advisor: Giovanni Ventorim Comarela.

Structural bioinformatics is dedicated to the study of three-dimensional structures of proteins and macromolecules. In this work, the interest is in protein structures. The availability of new sequences and protein structures in public databases has been occurring at a very fast pace, also increasing the need for automatic and efficient methods for extracting and understanding this large volume of data. According to Gao et al. [2018], bioinformatics is a science of mining and interpreting biological data. For them, the continuous and increasing flow of this data, as well as the need to address increasingly complex biomedical problems, has created challenging opportunities for data mining and machine learning researchers. Several strategies for structural protein classification have been proposed in recent years using sequence and structure based descriptors. In this research, evaluated the possibility of structural protein classification using unsupervised methods associated with successfully proposed characteristics in a well established structural classifier. Experiments were performed using 5 clustering algorithms from 4 different paradigms. The quality of the clusters was evaluated by the Silhouette Coefficient and the predicted labels were compared to the CATH database superfamily classifications using the Fowlkes Mallows Index and the verification of clusters homogeneity and completeness. The results show the unfeasibility of class level classification, since the rates achieved with Fowlkes Mollows did not reach 60%. On the other hand, they indicate a

considerable ability to classify at the superfamily level - an Index of over 70% was achieved with the Complete-Link method in the general clustering. The results are even more interesting when restricting the number of clusters, reaching an index of 78.5% for topologies with up to 25 superfamilies and 82.8% for topologies with up to 5 superfamilies. If still considered, clusters with an index equal to or greater than 85%, they represent approximately 40% of the topologies used, and of this group, almost half of the clusterings (48.19%) obtained a 100% similarity index, that is, in about 20% of the topologies, all proteins were clustered correctly.

Capítulo 1

INTRODUÇÃO

1.1 O problema e sua importância

A bioinformática é uma ciência multidisciplinar que aborda problemas de origem biológica por meio de recursos computacionais. A bioinformática estrutural envolve a representação, o armazenamento, a recuperação, análise e visualização de informações sobre as estruturas tridimensionais de proteínas [Gu & Bourne, 2011].

Proteínas são elementos fundamentais para os processos biológicos, atuando como catalisadoras de reações químicas e na composição de estruturas supramoleculares das células [Williams & Daviter, 2013]. Comparar estruturas proteicas é um processo que pode fornecer muita informação sobre sua evolução [Nelson & Cox, 2014]. Neste sentido, explorar diferentes estruturas contribui para uma melhor compreensão de seu funcionamento e potencializa a capacidade em lidar com certos tipos de doenças e questões relacionadas ao envelhecimento [Thornton, 2018].

A quantidade de novas estruturas disponíveis no *Protein Data Bank* (PDB) tem aumentado significativamente. Neste contexto, métodos automáticos e eficientes são cada vez mais necessários para explorar tais estruturas [Orengo et al., 1997]. Para Swindells et al. [1998], modelos capazes de identificar similaridades entre estruturas proteicas tornaram-se um assunto de grande relevância para pesquisa. Como resultado dessas transformações, bases de classificação estrutural, como CATH, têm buscado por melhorias contínuas através de mudanças em seus protocolos de classificação [Greene & Lewis, 2007]. Isso reforça, ainda mais, a necessidade de estratégias que contribuam para novas estruturas.

A motivação para o desenvolvimento deste trabalho surgiu de duas questões principais:

1. É possível gerar bons agrupamentos a partir da matriz de distâncias gerada pelo algoritmo da CSM?

2. Os grupos formados são similares à classificação CATH?

Dessa forma, foi realizada a classificação estrutural de proteínas por meio de uma abordagem não supervisionada, que utiliza como entrada os dados da *Cutoff Scanning Matrix* (CSM), uma matriz que representa os padrões de distância entre resíduos [Pires et al., 2011, 2013]. A utilização de tais dados justifica-se pelo fato de que proteínas estruturalmente similares também podem ter distâncias semelhantes entre seus resíduos [Pires et al., 2011]. Conforme Lee et al. [2007], dados estruturais podem ser usados para a detecção de similaridade funcional entre proteínas. Em Pires et al. [2011], os padrões de distância da CSM são usados para classificação estrutural e predição de funções proteicas.

A tarefa de classificação realizada aqui está dividida em duas partes: a classificação no nível classe, que busca dividir as estruturas de acordo com as quatro classes do primeiro nível CATH; e no nível superfamília, que considera as superfamílias homólogas presentes em cada topologia. Em ambos os casos, procurou-se mostrar a qualidade dos grupos formados e avaliar se tais grupos correspondem à classificação CATH.

Acredita-se na relevância de uma abordagem não supervisionada por não se limitar a prever apenas rótulos conhecidos. Acredita-se também que a principal contribuição seja discutir essa abordagem como uma forma de classificação estrutural, apresentando novas informações, diferentes daquelas comumente apresentadas por meio de métodos supervisionados.

1.2 Objetivos

1.2.1 Objetivo geral

Projetar, implementar e avaliar modelos e algoritmos para classificação estrutural de proteínas por meio de técnicas de aprendizagem não supervisionada.

1.2.2 Objetivos específicos

a) Coletar todas as estruturas disponíveis no PDB, assim como os descritores de classificação disponíveis na base de classificação estrutural CATH.

b) Construir um banco de dados contendo informações estruturais das proteínas coletadas e a classificação de cada estrutura.

c) Projetar, implementar e avaliar um modelo de aprendizagem não supervisionada baseado em descritores bem sucedidos de aprendizagem supervisionada, que seja capaz de tratar grandes volumes de dados.

d) Comparar os resultados da estratégia proposta com a classificação CATH.

1.3 Organização do trabalho

Este trabalho está organizado da seguinte forma: o Capítulo 2 compreende o Referencial Teórico, onde são apresentados os conceitos fundamentais para a implementação da estratégia proposta, e os trabalhos relacionados; o Capítulo 3 inclui Materiais e Métodos, onde são descritos os passos realizados para a classificação estrutural por meio de aprendizagem não supervisionada, envolvendo desde a seleção e o pré-processamento de dados ao agrupamento e avaliação; os resultados experimentais são apresentados e discutidos no Capítulo 4; e por fim, o Capítulo 5 faz algumas considerações sobre os resultados e propostas de trabalhos futuros.

Capítulo 2

REFERENCIAL TEÓRICO

2.1 Estrutura da proteína

As proteínas são as macromoléculas biológicas mais abundantes nos organismos vivos. Elas constituem a maior fração celular, além da água, e desempenham diversas funções importantes, atuando em atividades catalisadoras de reações químicas, na formação de estruturas supramoleculares, como receptoras de sinal e no transporte de substâncias específicas para dentro ou fora da célula [Williams & Daviter, 2013].

Tratam-se de polímeros formados por resíduos de aminoácidos conectados linearmente por meio de ligações covalentes. Os aminoácidos são moléculas orgânicas cuja estrutura caracteriza-se pela presença de dois grupos: um grupo carboxila (COOH) e um grupo amina (NH_2), ambos ligados a um átomo de carbono (Cα), e este também ligado a um átomo de hidrogênio (H) e a uma cadeia lateral R [Buxbaum, 2015]. A estrutura básica dos aminoácidos é apresentada na Figura 2.1.

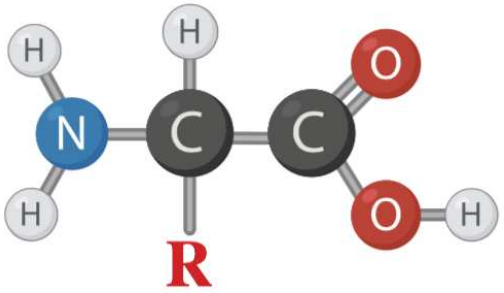


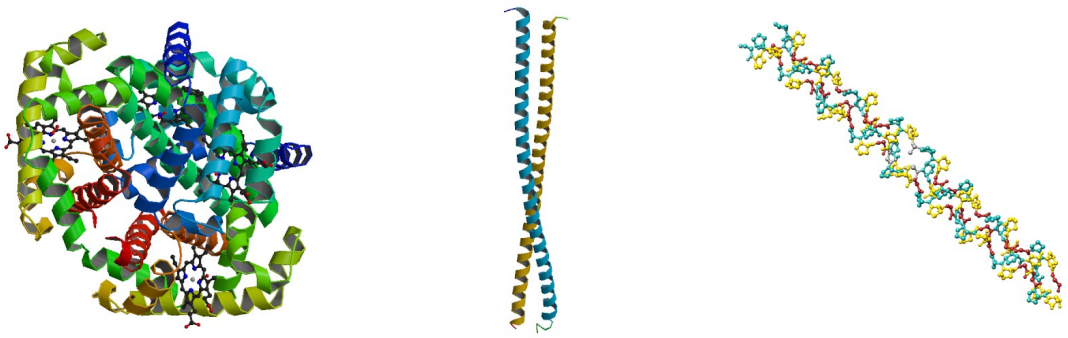
Figura 2.1. Estrutura básica de um aminoácido.

As proteínas são compostas basicamente por 20 principais aminoácidos, os quais se diferenciam pela cadeia lateral R [Fromm & Hargrove, 2012; Gu & Bourne, 2011]. A Tabela 2.1 apresenta o nome desses 20 aminoácidos, acompanhado pela sua abreviação e pelo seu símbolo.

Tabela 2.1. Lista dos 20 principais aminoácidos.

Aminoácido	Abreviação	Símbolo	Aminoácido	Abreviação	Símbolo
Alanina	ALA	A	Leucina	LEU	L
Arginina	ARG	R	Lisina	LYS	K
Asparagina	ASN	N	Metionina	MET	M
Asparato	ASP	D	Fenilalanina	PHE	F
Cisteina	CYS	C	Prolina	PRO	P
Glutamato	GLU	E	Serina	SER	S
Glutamina	GLN	Q	Treonina	THR	T
Glicina	GLY	G	Triptofano	TRP	W
Histidina	HIS	H	Tirosina	TYR	Y
Isoleucina	ILE	I	Valina	VAL	V

As funções desempenhadas pelas proteínas estão fortemente relacionadas com suas estruturas. A estrutura tridimensional de uma proteína resulta da combinação de diferentes fatores, como interações químicas presentes nessa proteína [Gu & Bourne, 2011]. A Figura 2.2 mostra as estruturas tridimensionais de três importantes proteínas - Hemoglobina, Queratina e Colágeno.



(a) (b) (c)

Figura 2.2. Exemplos de estruturas tridimensionais: (a) Hemoglobina Humana, obtida através da estrutura 2DN1 depositada no PDB. Esta proteína possui a função de armazenamento e transporte de oxigênio. (b) Queratina, obtida através da estrutura 1IC2. Ocorre em todos os vertebrados superiores, sendo o principal componente da rígida camada externa da epiderme. (c) Colágeno, obtido através da estrutura 1CAG. É a proteína mais abundante nos vertebrados, sendo um importante componente muscular.

A disposição espacial das proteínas resulta da sequência de aminoácidos que as compõem. Segundo Voet et al. [2014]; Marzzoco & Torres [2015], a configuração tridimensional de uma proteína, desde a sequência de seus aminoácidos à associação entre diferentes cadeias, pode ser descrita considerando quatro níveis de organização estrutural. Tais níveis são ilustrados na Figura 2.3 e descritos em seguida.

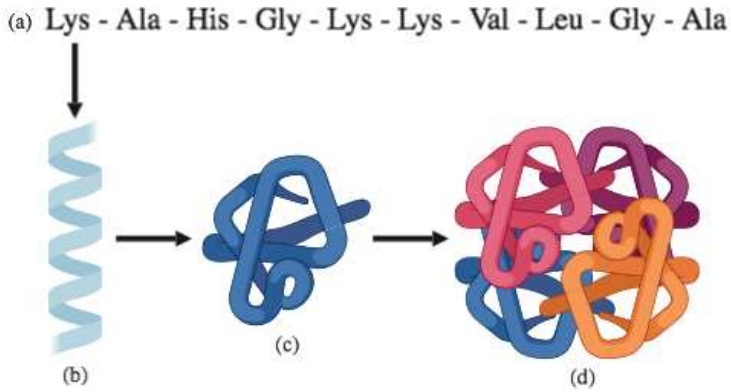


Figura 2.3. Níveis de organização estrutural das proteínas: (a) Estrutura primária, correspondente à sequência de aminoácidos na cadeia polipeptídica. (b) Estrutura secundária, correspondente ao arranjo estrutural dos átomos. (c) Estrutura terciária, correspondente a uma cadeia completa de proteína. (d) Estrutura quaternária, correspondente ao conjunto de cadeias completas.

A **estrutura primária** corresponde à sequência de aminoácidos da cadeia polipeptídica. Por convenção, esta sequência da-se na direção amino terminal $\Rightarrow$ carboxila terminal.

A **estrutura secundária** corresponde ao arranjo estrutural dos átomos, desconsiderando a conformação de suas cadeias laterais. Aqui, duas organizações da cadeia proteica são particularmente estáveis: o enrolamento da cadeia em torno de um eixo e a interação lateral entre os segmentos de uma mesma cadeia ou de cadeias distintas. Tais organizações são denominadas respectivamente, α-hélice e folha-β.

- α-hélice é um tipo de organização onde a espinha dorsal da proteína se enrola como um parafuso, e as cadeias laterais dos aminoácidos se projetam para fora da hélice.

- folha-β é um tipo de organização que integra pares de cadeias expandidas de aminoácidos. Sua estabilidade da-se por ligações de hidrogênio entre átomos de oxigênio no grupo CO e átomos de hidrogênio dos grupos NH em outra cadeia.

A **estrutura terciária** corresponde à estrutura tridimensional de um polipeptídeo completo, incluindo suas cadeias laterais. Envolve o arranjo tridimensional da conformação dos átomos da proteína, descrevendo todos os aspectos do enovelamento tridimensional do polipeptídeo.

A **estrutura quaternária** corresponde ao arranjo espacial de suas cadeias polipeptídicas (ou subunidades).

2.2 *Cutoff Scanning Matrix*

De acordo com Pires et al. [2011], a *Cutoff Scanning Matrix* (CSM) é uma técnica que produz uma matriz contendo as distâncias entre diferentes resíduos de proteínas. Cada linha da matriz representa a estrutura de uma proteína e cada coluna representa a frequência dos pares de resíduos a uma certa distância. Tal frequência corresponde ao número de contatos na proteína para aquela distância (Figura 2.4).

Figura 2.4. Representação de uma matriz de distâncias para duas estruturas coletadas aleatoriamente no PDB: (a) 1A5C e (b) 2QUH. Para cada estrutura, são apresentadas as frequências dos pares de resíduos no intervalo de distâncias de 0.0 Å a 5.0 Å, com passos de 0.2 Å. Ao todo, a matriz é composta por 26 colunas.

A representação da matriz de distâncias é um importante recurso para a análise estrutural de proteínas, fornecendo informações relevantes sobre sua conformação: proteínas com diferentes dobras e funções podem apresentar diferenças significativas na distribuição das distâncias entre os resíduos. Por outro lado, proteínas estruturalmente similares podem também apresentar distribuições semelhantes.

Ainda em Pires et al. [2011], os autores descrevem diferenças nas distribuições de densidade vetorial (frequência de contatos) para proteínas com classificações estruturais distintas (Figura 2.5). Nota-se que as diferenças entre as distribuições ocorrem em diferentes pontos de corte. Segundo os autores, tais variações não estão relacionadas apenas à composição estrutural secundária das proteínas, mas também ao seu empacotamento, incorporando informações importantes referentes à configuração espacial de cada estrutura.

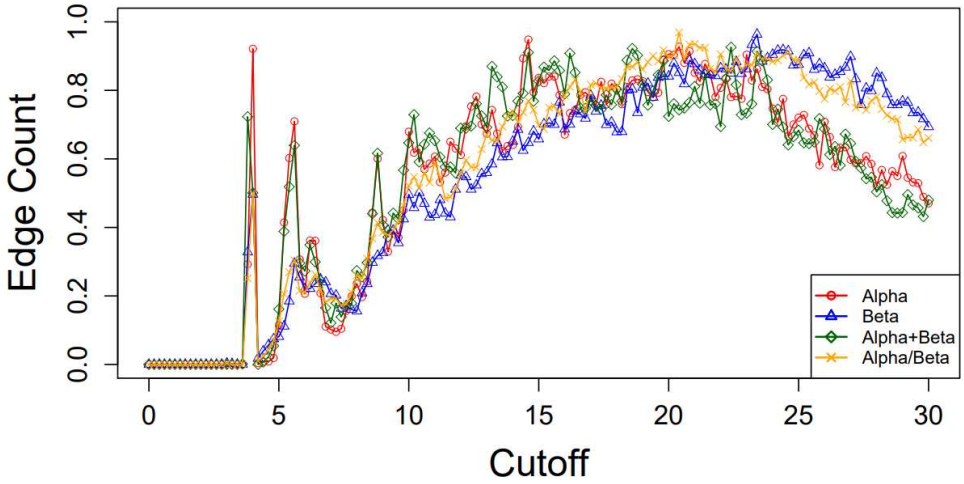


Figura 2.5. Distribuição de densidade vetorial para proteínas das classes de SCOP. Cada curva representa os valores médios de dez representantes selecionados aleatoriamente por classe. Figura extraída de Pires et al. [2011].

Para construir a matriz, calcula-se a distância euclidiana entre todos os Cα (outros centroides também podem ser escolhidos, como o Cβ), define-se um intervalo de distâncias e um passo de distância, e por fim, calcula-se a frequência dos pares de resíduos para cada distância dentro do intervalo definido (Algoritmo 1).

Algoritmo 1: Construção da CSM [Pires et al., 2011]

Require: $Proteinas$, CSM, $Dist_{MIN}$, $Dist_{MAX}$, $Dist_{INTERVAL}$

1: **for all** $proteina\ i \in (Proteinas)$ **do**
2: $j \leftarrow 0$
3: Calcula a distância entre todos os pares de $C\alpha$
4: **for** $dist \leftarrow Dist_{MIN}$ **to** $Dist_{MAX}$ **step** $Dist_{INTERVAL}$ **do**
5: $CSM[i][j] \leftarrow$ frequência dos pares de $C\alpha$ a uma distância $dist$
6: $j++$
7: **end for**
8: **end for**
9: **return** CSM

2.3 Base de classificação estrutural CATH

CATH é uma classificação hierárquica de domínios de proteínas divididos em quatro níveis: **Classe**, **Arquitetura**, **Topologia** e **Homologia** (C-A-T-H), onde as estruturas proteicas são decompostas em um ou mais domínios e classificadas em superfamílias homólogas [Cuff et al., 2009; Pearl et al., 2003, 2001]. A classificação CATH surgiu na década de 1990, baseada principalmente em avaliações manuais de semelhanças estruturais [Sillitoe et al., 2015]. Os dados são armazenados em uma base hierárquica, na qual cada estrutura é indexada com um número CATH específico, composto por quatro partes, e cada parte representa um nível dentro da hierarquia de classificação [Orengo et al., 1997].

A classificação no nível classe considera apenas o conteúdo estrutural de cada domínio. Em cada classe, os domínios são divididos por arquitetura, ou seja, de acordo com as semelhanças entre suas conformações. As arquiteturas são subdivididas em topologias ou dobras, onde são considerados os contatos entre estruturas secundárias. Por fim, os domínios em cada topologia são divididos por superfamília homóloga, e compartilham semelhanças sequenciais, estruturais e funcionais [Cuff et al., 2009; Pearl et al., 2003, 2001]. Proteínas com uma mesma classificação CATH possuem estruturas globalmente semelhantes, embora suas sequências possam ser muito diferentes [Swindells et al., 1998].

O nível classe da hierarquia CATH agrupa as estruturas de acordo com o conteúdo da estrutura secundária [Knudsen & Wiuf, 2010]. A estrutura secundária pode ser entendida por determinadas formas estruturais compartilhadas por muitas proteínas, onde, como citado anteriormente, destacam-se as estruturas α-hélice e folha-β. Ainda no primeiro nível de classificação, CATH inclui estruturas com conteúdo estrutural muito baixo - *Few Secondary Structures* [Orengo et al., 1997].

A Tabela 2.2 apresenta a composição de cada classe, considerando a quantidade de arquiteturas, topologias, superfamílias homólogas (A-T-H) e domínios.

Tabela 2.2. Composição das classes do primeiro nível da hierarquia CATH.

	C	A	T	H	Domínios
1	*Mainly Alpha*	5	405	2.174	90.302
2	*Mainly Beta*	21	244	1.395	110.260
3	*Alpha Beta*	14	634	2.428	229.776
4	*Few Secondary Structures*	1	108	122	4.519

2.4 Mineração de dados e a descoberta de conhecimento

A mineração de dados é o processo responsável pela extração de conhecimento em grandes volumes de dados. Trata-se de uma disciplina inter e multidisciplinar, envolvendo conhecimentos de diferentes áreas como banco de dados, estatística, reconhecimento de padrões, visualização de dados, recuperação de informação, inteligência artificial, entre outras [de Castro & Ferrari, 2016].

Para Fayyad et al. [1996]; Cios et al. [2007], a mineração de dados integra um processo mais amplo, conhecido como descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases* - KDD), representado na Figura 2.6.

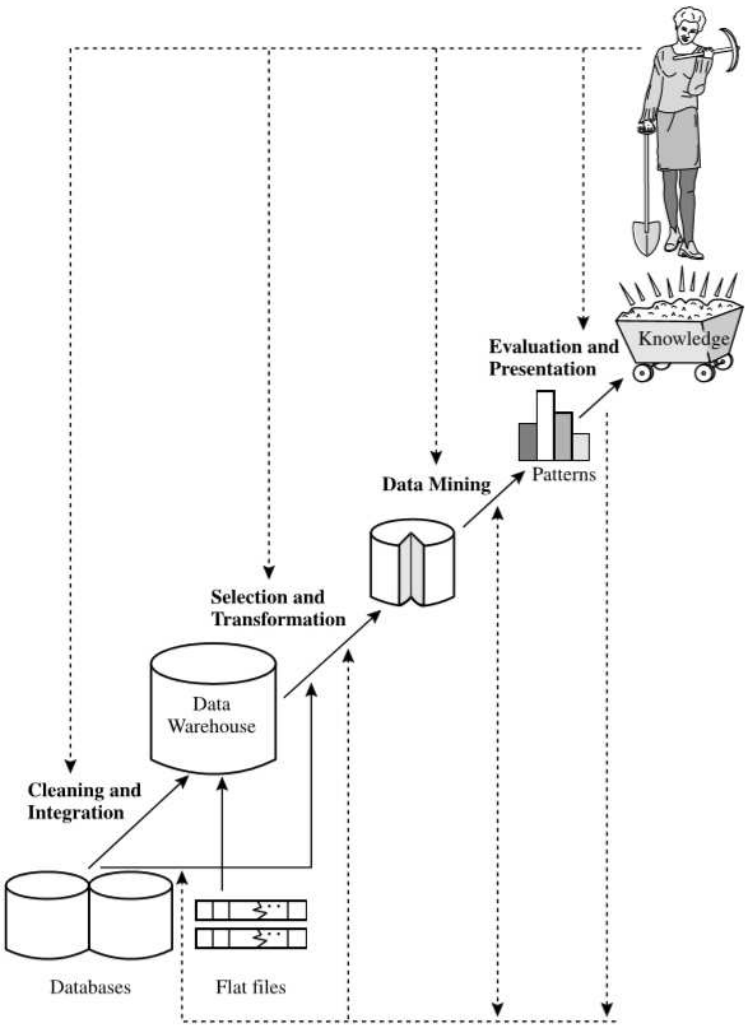


Figura 2.6. Mineração de dados como uma etapa no processo de KDD. Figura extraída de Han & Kamber [2006].

De acordo com Han & Kamber [2006], a descoberta de conhecimento em bases de dados consiste em uma sequência de atividades:

- Na **limpeza e integração** são removidos ruídos e dados inconsistentes. Nesta etapa ainda podem ser combinadas diferentes fontes de dados.
- Na **seleção e transformação** os dados mais relevantes são selecionados e preparados para a tarefa de mineração. Envolve técnicas como normalização (tem como objetivo alcançar maior precisão e eficiência dos algoritmos, em especial aqueles que envolvem medições de distância) e redução de dimensionalidade (tem como objetivo eliminar recursos redundantes e melhorar o desempenho).
- A **mineração de dados** é o processo essencial na descoberta de conhecimento, onde diversas técnicas podem ser aplicadas na busca por padrões.
- Na **avaliação e apresentação** ocorre a verificação dos padrões encontrados e apresentação ao usuário por meio de técnicas de visualização e representação do conhecimento.

2.4.1 Decomposição em valores singulares

A redução de dimensionalidade permite obter a representação reduzida de um conjunto de dados com resultados bem próximos do conjunto original. É um processo que contribui para diminuir ruídos e melhorar o desempenho computacional durante a execução dos algoritmos de mineração [Han & Kamber, 2006].

A decomposição em valores singulares (*Singular Value Decomposition* - SVD) é uma técnica para redução de dimensionalidade que possui aplicações em diferentes áreas. Para Trefethen & III [1997], trata-se da fatoração de uma matriz qualquer em três outras matrizes com características importantes. Segundo Elden [2007], a SVD decompõe uma matriz na seguinte forma:

$$A = U\Sigma V^T \tag{2.1}$$

Onde U é uma matriz ortogonal de dimensões $m \times m$ e suas colunas são autovetores de AA^T, V é uma matriz ortogonal de dimensões $n \times n$ e suas colunas são autovetores de A^TA, e Σ é uma matriz diagonal de dimensões $m \times n$ composta por valores reais não negativos denominados **valores singulares**, sendo estes ordenados dos mais significativos para os menos significativos. A SVD também pode ser representada na forma da Equação 2.2, onde u_i e v_i representam os i-ésimos autovetores de U e V respectivamente e σ_i é o i-ésimo valor singular de Σ.

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T \tag{2.2}$$

Para Pires et al. [2011], ao considerar um subconjunto específico de valores singulares de tamanho $k < n$, onde n é o posto de A, é possível obter uma matriz A_k aproximada da original A:

$$A \approx A_k = U_k \Sigma_k V_k^T \tag{2.3}$$

Ainda segundo os autores, ao representar um conjunto de dados com uma quantidade reduzida de valores singulares, há uma tendência no agrupamento de alguns itens que não seriam agrupados no conjunto completo. A Figura 2.7 ilustra a redução de dimensionalidade com SVD.

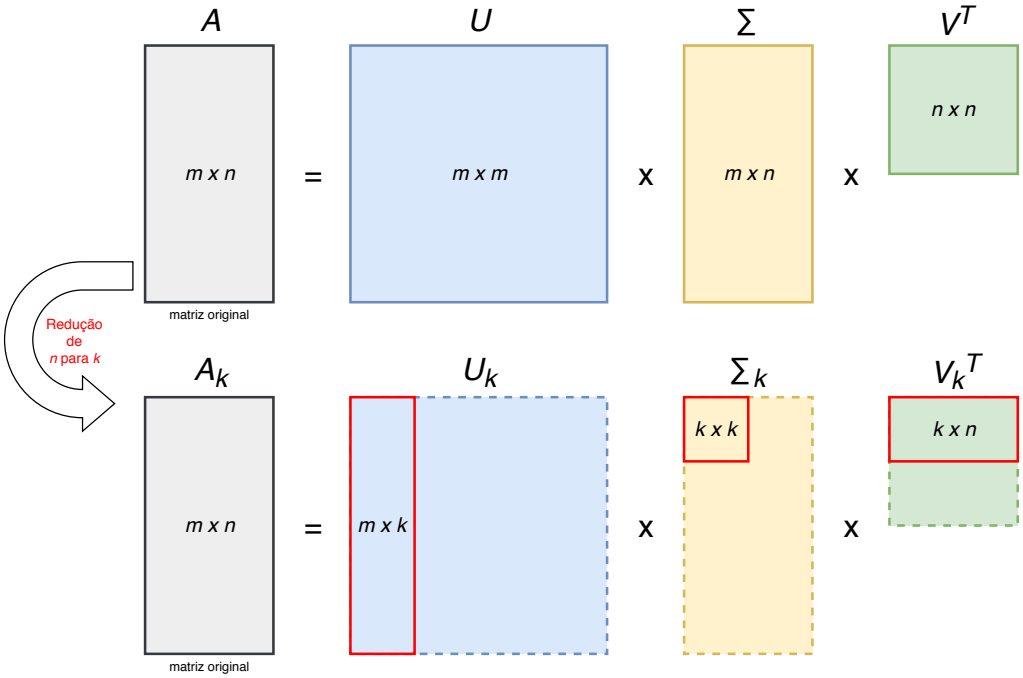


Figura 2.7. Representação da SVD em uma redução de n para k dimensões. Figura adaptada de Elden [2007].

2.4.2 Análise de agrupamento

Em mineração de dados o agrupamento é um exemplo de aprendizagem não supervisionada, onde não há classes pré-definidas e treinamentos rotulados [Han & Kamber, 2006]. Dessa forma, a principal diferença para a aprendizagem supervisi-

onada está nos dados de entrada, que não possuem rótulos. Podemos dizer que a aprendizagem não supervisionada busca simular a aprendizagem humana, mas sem supervisão explícita [Pacheco, 2015]. Para Fisher et al. [2014], em métodos não supervisionados não são fornecidas orientações, eles devem descobrir categorias úteis de dados com base em heurísticas internas. Segundo Mishra & Rath [2018], esse tipo de abordagem tem como principal objetivo encontrar centros de grupos próximos e determinar os melhores agrupamentos.

Análise de agrupamento é uma tarefa que separa objetos considerando apenas as informações que os descrevem e seus relacionamentos. O objetivo é maximizar a similaridade intra-grupo e minimizar a similaridade inter-grupo; ou seja, objetos de um mesmo grupo devem ser similares entre si e diferentes de objetos em outros grupos. Quanto maior a homogeneidade dentro de cada grupo e a diferença entre os grupos, melhor a qualidade do agrupamento [Tan et al., 2005].

Um grupo corresponde a uma coleção de objetos semelhantes entre si e diferentes de objetos em outros grupos. A escolha do algoritmo de agrupamento está relacionada aos tipos de dados disponíveis e ao objetivo específico da aplicação [Han & Kamber, 2006]. A seguir são descritas as técnicas de agrupamento utilizadas nesse trabalho.

2.4.2.1 K-*means*

De acordo com Han & Kamber [2006], trata-se de um método de agrupamento particional, com o qual, dado um conjunto de dados, são construídas k partições, onde cada partição representa um grupo. Os dados são classificados em k grupos que juntos satisfazem os seguintes requisitos:

- Em cada grupo deve haver pelo menos um objeto;
- Cada objeto deve pertencer a exatamente um grupo.

O algoritmo K-*means* recebe um parâmetro de entrada k e particiona um conjunto de n objetos em k grupos, de maneira que haja alta similaridade entre objetos do mesmo grupo e baixa similaridade entre objetos de grupos distintos. A similaridade em cada grupo é medida em relação ao valor médio dos objetos daquele grupo - o que denomina-se **centroide**. Este processo funciona da seguinte forma: inicialmente são selecionados aleatoriamente k objetos como sendo os centros de grupo; cada objeto restante é então atribuído ao grupo mais próximo, de acordo com a distância entre este objeto e o centroide do grupo; em seguida, a média de cada grupo é recalculada. Tal processo se repete até que não haja mais alterações

nos centroides já calculados. O Algoritmo 2 apresenta um pseudocódigo de alto nível para o método K-*means*.

Algoritmo 2: Método K-*means* [Tan et al., 2005]

Require: *DataSet*, k

1: Seleciona k pontos como centroides iniciais
2: **repeat**
3: Forma k grupos atribuindo cada ponto ao seu centroide mais próximo
4: Recalcula o centroide de cada grupo
5: **until** Os centroides não mudarem

Para Tan et al. [2005], K-*means* é uma técnica de agrupamento baseada em protótipo, que tenta encontrar um número de k grupos definido pelo usuário, sendo cada grupo representado pelo seu centroide. Para atribuir um objeto ao centroide mais próximo é preciso utilizar alguma medida de proximidade. Em conjuntos de dados onde esta medida é a distância Euclidiana, podemos aplicar a soma do erro quadrado (*sum of the squared error* - SSE) ou dispersão, que utilizando a notação da Tabela 2.3, pode ser definida como:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)^2 \tag{2.4}$$

Utilizando ainda a notação da Tabela 2.3, o centroide do i-ésimo grupo pode ser definido como:

$$c_i = \frac{1}{m_i} \sum_{x \in C_i} x \tag{2.5}$$

Tabela 2.3. Notação utilizada para calcular SSE e c_i

Símbolo	Descrição
x	Um objeto do conjunto de dados
C_i	O i-ésimo grupo
c_i	Centroide do grupo C_i
c	Centroide de todos os pontos
m_i	Número de objetos presentes no i-ésimo grupo
m	Número de objetos no conjunto de dados
K	Número de grupos
$dist$	Distância entre dois objetos no espaço Euclidiano

A Figura 2.8 ilustra o agrupamento de dados baseado no algoritmo K-*means*, de acordo com o Algoritmo 2.

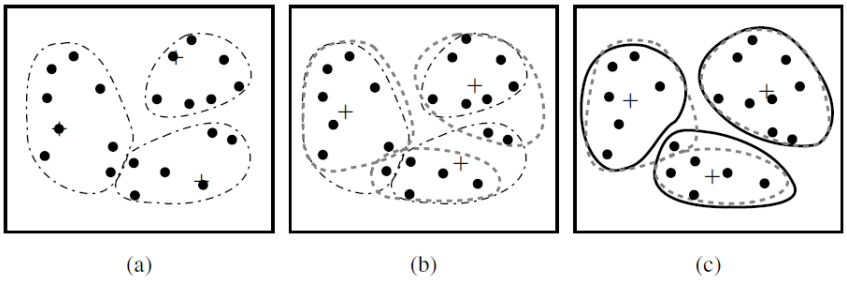


Figura 2.8. Processo de agrupamento de dados com K-*means*: (a) Inicialmente são selecionados k pontos como sendo os centros de grupos - os chamados centroides, representados pelo símbolo "+"; cada objeto é então distribuído para um grupo com base no centroide mais próximo. (b) Em seguida, o valor médio de cada grupo é recalculado com base nos objetos atuais, ou seja, os centroides são atualizados. Dessa forma, os objetos são novamente distribuídos com base nos novos centroides. Essa redistribuição forma silhuetas circundadas com curvas tracejadas. (c) O processo de reatribuição iterativa, no qual cada objeto vai sendo colocado em seu grupo correspondente, ocorre até que nenhuma redistribuição dos objetos em qualquer grupo aconteça. Figura extraída de Han & Kamber [2006].

2.4.2.2 Métodos hierárquicos

Conforme Han & Kamber [2006], os métodos de agrupamento hierárquicos dividem os objetos de dados em uma árvore de grupos. Eles podem ser classificados como **aglomerativos** ou **divisivos**, de acordo com a sua decomposição hierárquica - de baixo para cima (fusão) ou de cima para baixo (divisão).

A abordagem aglomerativa inicia com cada objeto formando grupos isolados. Tais objetos ou grupos mais próximos vão sendo mesclados sucessivamente, até que todos os grupos estejam fundidos em um nível superior da hierarquia ou até que uma condição de parada seja alcançada. Por outro lado, a abordagem divisiva inicia com todos os objetos em um mesmo grupo. A cada iteração, cada grupo é dividido em grupos menores até que um objeto forme um grupo ou que uma condição de parada seja alcançada.

Para Tan et al. [2005], as técnicas de agrupamento hierárquico são a segunda categoria mais importante em análise de agrupamento. Assim como K-*means*, estas abordagens já estão bastante consolidadas. Neste trabalho, entre os métodos de agrupamento utilizados, dois são hierárquicos aglomerativos. O Algoritmo 3 apresenta um pseudocódigo de alto nível para esta categoria.

Algoritmo 3: Método Hierárquico Aglomerativo [Tan et al., 2005]

1: Calcula a matriz de proximidade, se necessário
2: **repeat**
3: Mescla os dois conjuntos mais próximos
4: Atualiza a matriz de proximidade de modo a refletir a proximidade entre o novo grupo e os grupos originais
5: **until** Permanecer apenas um grupo

Para Tan et al. [2005], a definição de proximidade de grupo é o que distingue as diferentes técnicas hierárquicas aglomerativas - **MIN** define a proximidade do grupo como a distância entre os dois pontos mais próximos em grupos distintos; **MAX** define a proximidade do grupo como a distância entre os dois pontos mais distantes em grupos distintos; por fim, a **média do grupo** define a proximidade como a distância média de todos os pares de pontos em grupos distintos. A Figura 2.9 mostra as diferentes definições de proximidade de grupo para o agrupamento hierárquico, onde são apresentadas respectivamente as distâncias MIN, MAX e média.

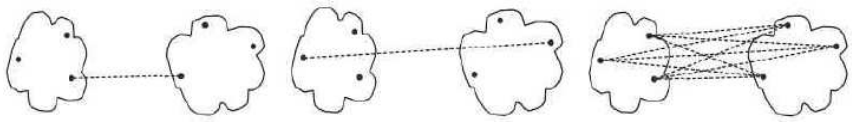


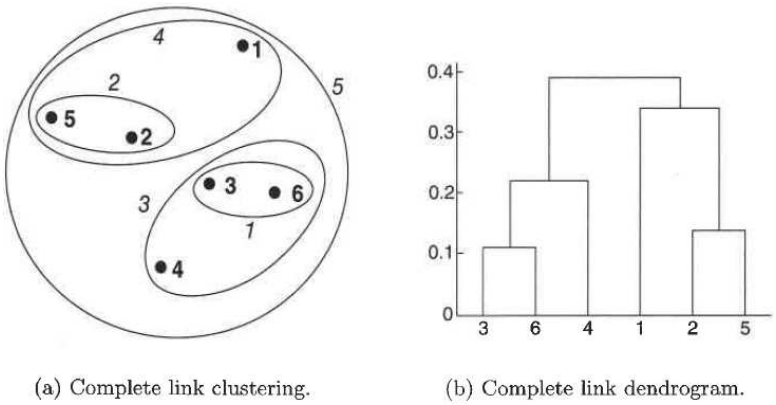
Figura 2.9. A sequência apresenta respectivamente as definições de proximidade MIN (distância entre os dois pontos mais próximos em grupos diferentes), MAX (distância entre os dois pontos mais distantes em grupos diferentes) e média de grupo (distância média de todos os pares de pontos em grupos diferentes). Figura extraída de Tan et al. [2005].

A seguir, apresentamos uma breve descrição dos métodos hierárquicos aglomerativos aplicados neste trabalho:

- ***Complete-Link***: utiliza MAX como proximidade de grupo; ou seja, a proximidade entre dois grupos é definida como a maior distância entre dois pontos de grupos distintos. Trata-se de um método menos suscetível a ruídos e *outliers* [Tan et al., 2005].

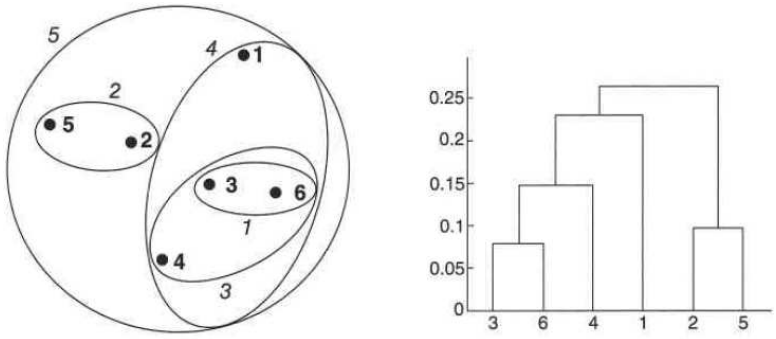
- ***Ward***: utiliza como proximidade de grupo o aumento no erro quadrado, resultante da unificação de dois grupos. Assim como K-*means*, utiliza a função objetivo SSE [Tan et al., 2005].

As Figuras 2.10 e 2.11 ilustram diferentes representações de um agrupamento hierárquico aglomerativo com os métodos *Complete-Link* e *Ward* respectivamente.



(a) Complete link clustering. (b) Complete link dendrogram.

Figura 2.10. Diferentes representações para um agrupamento hierárquico aglomerativo com *Complete-Link*. Figura extraída de Tan et al. [2005].



(a) Ward's clustering. (b) Ward's dendrogram.

Figura 2.11. Diferentes representações para um agrupamento hierárquico aglomerativo com *Ward*. Figura extraída de Tan et al. [2005].

2.4.2.3 DBSCAN

Grande parte dos métodos de particionamento realiza o agrupamento com base na distância entre objetos. Tais métodos podem encontrar apenas grupos esféricos. Para suprir as dificuldades em encontrar grupos de forma arbitrária, sem uma definição inicial do número de grupos, foram desenvolvidos os métodos baseados no conceito de densidade [Han & Kamber, 2006].

Esse tipo de agrupamento encontra regiões de alta densidade separadas por regiões de baixa densidade. DBSCAN é um sistema simples e eficaz baseado em densidade. Trata-se de uma estratégia em que o número de grupos é automaticamente determinado pelo algoritmo com base em uma distância de corte denominada EPS [Tan et al., 2005]. O Algoritmo 4 apresenta um pseudocódigo de alto nível para o método DBSCAN. Em seguinda, a Figura 2.12 ilustra dois exemplos de agrupamentos com este método para valores distintos de EPS.

Algoritmo 4: Método DBSCAN [Tan et al., 2005]

1: Rotula todos os pontos como pontos centrais, de borda ou de ruído
2: Elimina pontos de ruído
3: Inclui uma borda entre todos os pontos centrais que estão dentro de uma certa distância (EPS)
4: Transforma grupos de pontos centrais conectados em um grupo separado
5: Atribui cada ponto de fronteira a um dos grupos de seus pontos

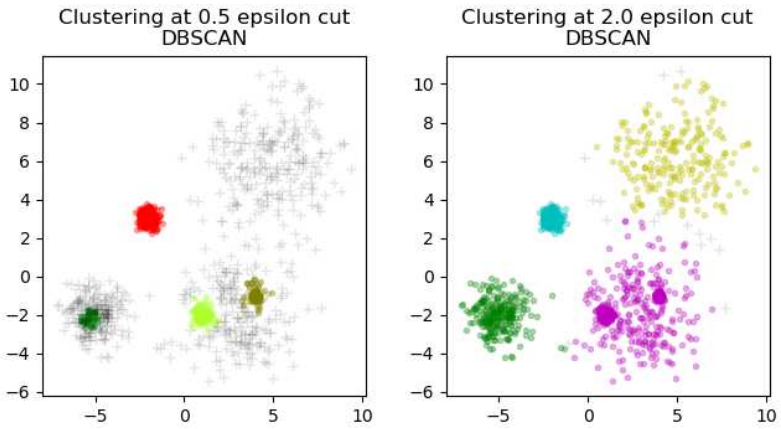


Figura 2.12. Exemplos de agrupamentos com DBSCAN para valores distintos de EPS. Figura extraída de scikit learn [2019].

2.4.2.4 Agrupamento Espectral

De acordo com Bach & Jordan [2003]; von Luxburg [2007], a abordagem espectral de agrupamento inclui um conjunto de técnicas que particionam os dados por meio de autovetores de matrizes, como a matriz de similaridade. Para von Luxburg [2007], entre os métodos modernos de agrupamento, este é hoje um dos mais populares por apresentar vantagens importantes em relação aos algoritmos tradicionais - ao contrário do algoritmo K-*means*, por exemplo, o método espectral não faz nenhuma suposição relacionada à forma do conjunto de dados. Para Han & Kamber [2006], isso lhe confere a capacidade de manipular dados altamente não-lineares e suas fortes conexões com a geometria diferencial o tornam capaz de descobrir a estrutura múltipla do espaço. No entanto, uma desvantagem deste método é a necessidade de usar todos os pontos para a incorporação, o que o torna computacionalmente caro para grandes conjuntos de dados.

O método primeiro realiza a incorporação espectral (redução de dimensionalidade) nos dados originais, aplicando em seguida algum algoritmo de agrupamento tradicional, como o K-*means* [Han & Kamber, 2006]. Dois objetos matemáticos são usados para o agrupamento espectral: grafos de similaridade e matrizes La-

placianas. A partir do grafo de similaridade, calcula-se a matriz Laplaciana, seus autovalores e autovetores, tomando-se os primeiros k autovetores como dimensões na nova representação do conjunto de dados. A principal vantagem deste método está na alteração da representação dos objetos x_i no espaço original para pontos y_i no espaço km-dimensional. Tal alteração aumenta as propriedades de grupo presentes nos dados, de maneira que grupos possam ser identificados facilmente na nova representação, por meio de um método simples, como K-*means*, após a redução de dimensionalidade. Um exemplo de implementação do método espectral é mostrado no Algoritmo 5, aplicado quando utiliza-se a matriz Laplaciana normalizada assimétrica L_{rw} [von Luxburg, 2007]. Em seguinda, a Figura 2.13 ilustra um exemplo de agrupamento com este método.

Algoritmo 5: Agrupamento Espectral [von Luxburg, 2007]

Require: k grupos, matriz de similaridades $S \in R^{N \times N}$

1: Construir um grafo de similaridades e definir A como sua matriz de adjacências
2: Calcular a matriz Laplaciana não normalizada L
3: Calcular a matriz Laplaciana normalizada assimétrica L_{rw}
4: Calcular os primeiros k autovetores $u_i, ..., u_k$ de L_{rw}
5: Seja $U \in R^{N \times K}$ a matriz contendo os autovetores $u_i, ..., u_k$ como colunas
6: **for all** $i = 1...N$ **do**
7: Seja $y_i \in R^k$ o vetor correspondente à i-ésima linha de U
8: **end for**
9: Agrupar os pontos $(y_i)_{i=1...N}$ em R^K com K-Means, gerando grupos $C_1, ..., C_k$
10: **return** grupos $A_1, ..., A_k$ com $A_i = \{j | y_j \in C_j\}$

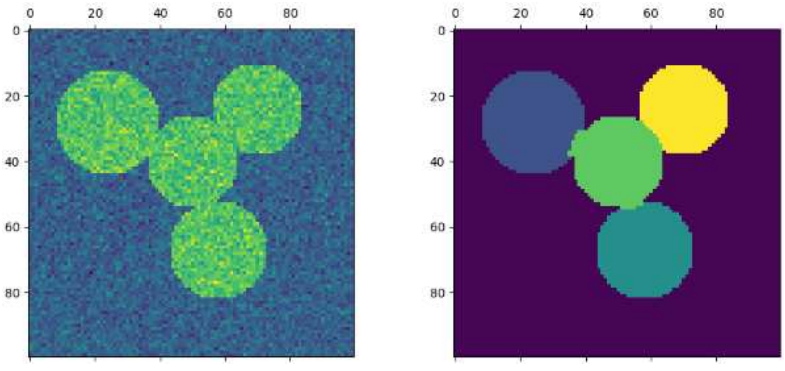


Figura 2.13. Exemplo de agrupamento espectral: a partir de uma imagem com círculos conectados (esquerda), o agrupamento espectral é aplicado para separá-los (direita). Figura extraída de scikit learn [2019].

Comparação entre os métodos de agrupamento A Figura 2.14 apresenta uma comparação entre os métodos de agrupamento descritos neste trabalho utilizando diferentes conjuntos de dados.

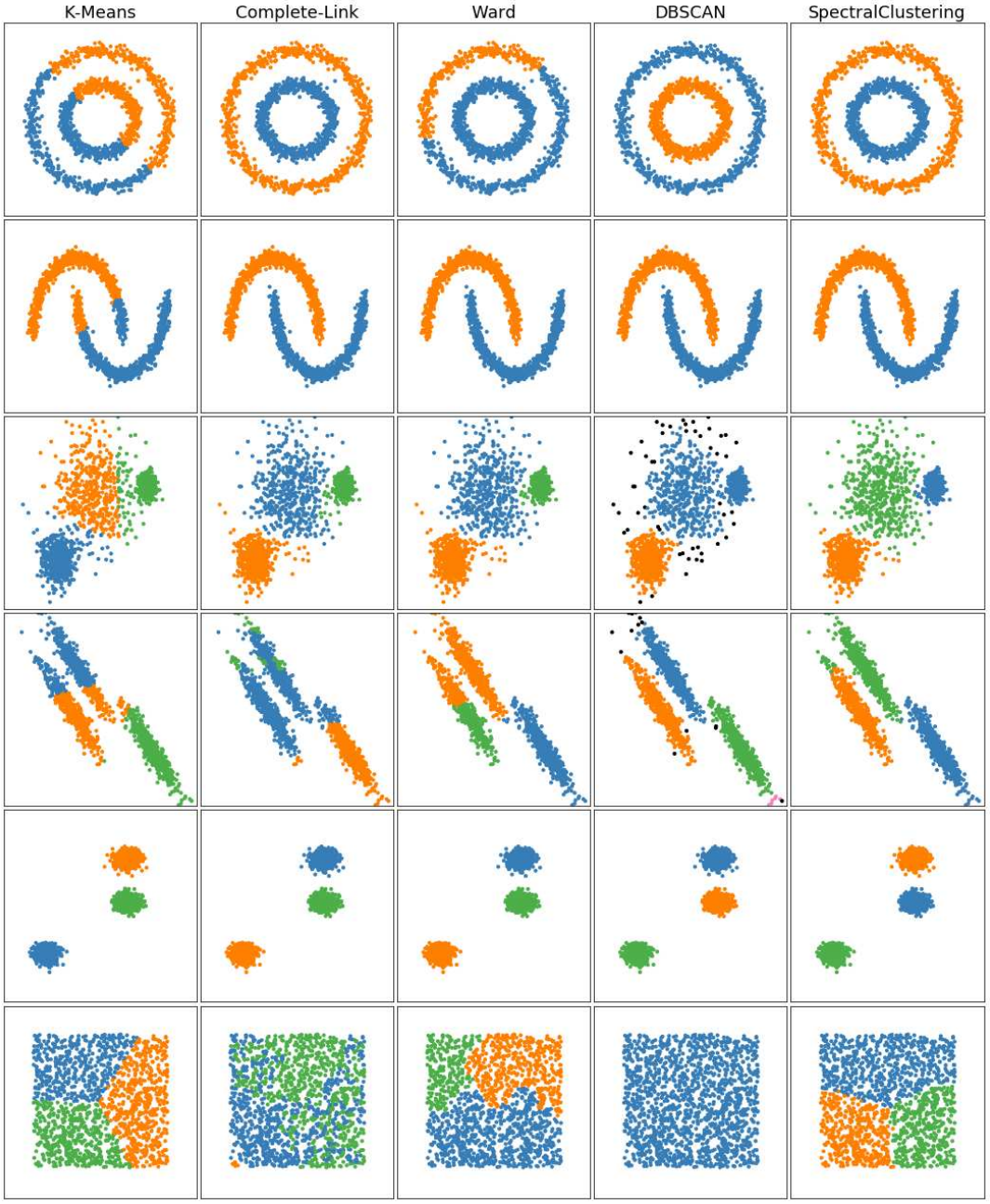


Figura 2.14. Comportamento dos métodos de agrupamento em diferentes conjuntos de dados: cada coluna representa um método e cada linha um conjunto de dados. Neste caso, o último conjunto (linha 6) é constituído por objetos homogêneos, não havendo uma separação adequada, ou seja, não há a formação de bons agrupamentos. Figura adaptada de scikit learn [2019].

2.4.3 Validação de grupos

O processo de validação permite avaliar o resultado produzido por um método de agrupamento. Para Tan et al. [2005], esta é uma importante etapa da aprendizagem não supervisionada que pode ser utilizada para:

- Determinar a tendência de um agrupamento, distinguindo conjuntos de dados aleatórios e não aleatórios;
- Determinar o número correto de grupos;
- Comparar rótulos de agrupamento com rótulos previamente conhecidos;
- Comparar grupos e determinar qual é o melhor.

De acordo com Jain & Dubes [1988]; Dziopa [2016], a validação pode ser classificada em três critérios:

1. **Interno**: avalia o quão boa é a estrutura de um agrupamento, desconsiderando informações externas.
2. **Externo**: avalia até que ponto rótulos de grupos produzidos por um método de agrupamento são equivalentes a rótulos de classes externas.
3. **Relativo**: compara dois grupos ou agrupamentos distintos.

A seguir, descrevemos as medidas de validação (interna e externa) aplicadas neste trabalho.

2.4.3.1 Coeficiente de Silhueta

O Coeficiente de Silhueta avalia a qualidade de um agrupamento com base na proximidade entre os objetos de um grupo e na proximidade desses objetos ao grupo mais próximo. Trata-se de uma medida de validação interna, usada para avaliar a formação dos melhores grupos quando não há rótulos conhecidos - rótulos observados antes do agrupamento [Rousseeuw, 1987; Dziopa, 2016]. Tal medida é definida para cada objeto e pode ser expressa pela seguinte equação:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{2.6}$$

Onde a_i é a distância média entre um objeto x_i e outros objetos do mesmo grupo e b_i a distância média entre um objeto x_i e outros objetos do grupo mais próximo.

A pontuação assumida por s_i pode variar de 1 a -1, sendo que quanto mais próxima de 1 melhor a alocação de x_i e quanto mais próxima de -1 pior a alocação. Pontuações próximas de zero podem indicar a formação de grupos sobrepostos.

- $s_i = 1$ indica que um objeto x_i está próximo aos objetos de seu grupo e distante dos demais grupos;
- $s_i = 0$ indica que um objeto x_i está próximo ao limite entre dois grupos vizinhos;
- $s_i = -1$ indica que um objeto x_i está mais próximo a outro grupo de que seu próprio grupo.

A Figura 2.15 ilustra os elementos envolvidos no cálculo de s_i, onde x_i pertence ao grupo A e pretende-se encontrar seu vizinho mais próximo entre B e C.

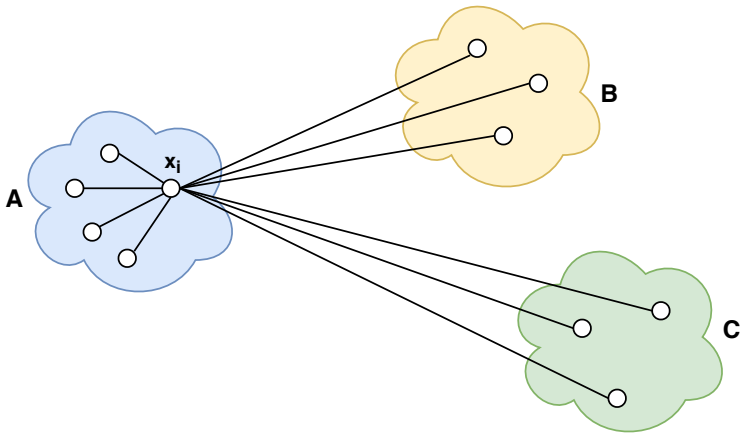


Figura 2.15. Elementos envolvidos no cálculo de s_i: considerando três grupos A, B e C, para qualquer objeto x_i alocado em A tem-se $m(x_i, A)$ a média das distâncias de x_i aos demais objetos de A, $m(x_i, B)$ a média das distâncias de x_i aos objetos de B e $m(x_i, C)$ a média das distâncias de x_i aos objetos de C. Após calcular todos os $m(x_i, B)$ para A≠B e todos os $m(x_i, C)$ para A≠C, seleciona-se os menores valores. Neste exemplo, o grupo B, para o qual o menor valor é obtido, é definido como vizinho de x_i. Através da iteração deste processo, obtém-se o grupo vizinho de cada um dos objetos x_i. Figura adaptada de Rousseeuw [1987].

O Coeficiente de Silhueta do agrupamento pode ser expresso pela equação 2.7, onde N representa o número de grupos envolvidos.

$$S = \frac{1}{N} \sum_{i=1}^{N} s_i \tag{2.7}$$

2.4.3.2 Índice *Fowlkes Mallows*

O índice *Fowlkes Mallows* (*Fowlkes Mallows Index* - FMI) é uma medida de validação externa que permite avaliar os grupos formados considerando rótulos de classificação já conhecidos (rótulos observados no mundo real). De acordo com Dziopa [2016], a pontuação FMI é definida como a média geométrica da **precisão** (*precision*) e **revocação** (*recall*). Nesse contexto, alguns conceitos são fundamentais para o entendimento da validação de agrupamento por meio de FMI:

- **Verdadeiro Positivo** (*True Positive* - TP): são os elementos classificados corretamente. Em análise de agrupamento, dado um conjunto X, corresponde aos pares de objetos $x_a, x_b \in X$ que estão num mesmo grupo no conjunto de rótulos observados e também no conjunto de rótulos previstos;
- **Falso Positivo** (*False Negative* - FN): pares de objetos $x_a, x_b \in X$ que estão em grupos distintos no conjunto de rótulos observados mas num mesmo grupo no conjunto de rótulos previstos;
- **Falso Negativo** (*False Positive* - FP): pares de objetos $x_a, x_b \in X$ que estão num mesmo grupo no conjunto de rótulos observados mas em grupos distintos no conjunto de rótulos previstos;
- **Verdadeiro Negativo** (*True Negative* - TN): pares de objetos $x_a, x_b \in X$ que estão em grupos distintos no conjunto de rótulos observados e também no conjunto de rótulos previstos.

A Figura 2.16 ilustra os conceitos de TP, FP, FN e TN por meio de uma matriz confusão:

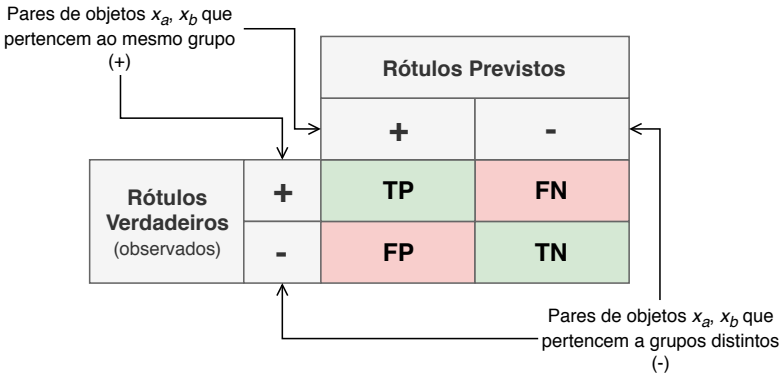


Figura 2.16. Matriz confusão para os conceitos de TP, FP, FN e TN em análise de agrupamento. Figura adaptada de Tan et al. [2005].

A partir dos conceitos descritos acima, pode-se definir as seguintes equações que nos levam ao FMI:

$$precision = \frac{TP}{TP + FP} \tag{2.8}$$

$$recall = \frac{TP}{TP + FN} \tag{2.9}$$

$$FMI = \frac{TP}{\sqrt{(TP + FP)(TP + FN)}} \tag{2.10}$$

A pontuação FMI pode variar de 0 a 1, onde valores próximos de 0 indicam atribuições de rótulos independentes e valores próximos de 1 indicam equivalência significativa. Valores exatamente iguais a 0 indicam atribuições puramente independentes e valores iguais a 1 indicam atribuições iguais [scikit learn, 2019].

2.4.3.3 Homogeneidade, Completude e Medida-V

Quando conhecemos as atribuições de classe verdadeiras (rótulos de classificação), é possível definir alguma métrica intuitiva a partir da análise de entropia condicional. Rosenberg & Hirschberg [2007] descrevem dois objetivos desejáveis para qualquer atribuição de grupo:

- **Homogeneidade**: cada grupo contém apenas membros de uma única classe.

$$h = 1 - \frac{H(C|K)}{H(C)} \tag{2.11}$$

- **Completude**: todos os membros de uma classe são atribuídos ao mesmo grupo.

$$c = 1 - \frac{H(K|C)}{H(K)} \tag{2.12}$$

Onde $H(C|K)$ é a entropia condicional das classes dadas as atribuições do grupo e $H(C)$ é a entropia das classes.

$$H(C|K) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{n_{c,k}}{n} \cdot log\left(\frac{n_{c,k}}{n_k}\right) \tag{2.13}$$

$$H(C) = -\sum_{c=1}^{|C|}\frac{n_c}{n} \cdot log\left(\frac{n_c}{n}\right) \tag{2.14}$$

Sendo C o número de classes, K o número de grupos, n o número geral de amostras, n_c e n_k o número de amostras da classe c e do grupo k respectivamente, e $n_{c,k}$ o número de amostras da classe c atribuído ao grupo k. A entropia condicional de grupos dada a classe $H(K|C)$ e a entropia de grupos $H(K)$ são definidas de forma similar.

As pontuações assumidas para Homogeneidade e Completude podem variar de 0 a 1, onde quanto mais próxima de 1 melhor a pontuação e quanto mais próxima de 0 pior a pontuação.

Rosenberg & Hirschberg [2007] ainda descrevem a **Medida-V** como a média harmônica da Homogeneidade e Completude, conforme a Equação 2.15.

$$v = 2 \cdot \frac{h \cdot c}{h + c} \tag{2.15}$$

2.5 Trabalhos relacionados

Nesta seção são apresentados de forma cronológica alguns trabalhos relacionados à classificação estrutural de proteínas e quais os principais pontos que os diferem deste trabalho.

Em Rogen & Fain [2003], os autores desenvolveram uma estratégia para análise e comparação de proteínas a partir de invariantes topológicos. Para isso, foi criada uma medida de similaridade denominada *Scaled Gauss metric* (GSM), aplicada na classificação automática de estruturas da base CATH. Tal estratégia alcançou 95,51% de atibuições corretas.

Em uma outra estratégia, Sun & Huang [2006] aplicaram *Support Vector Machines* (SVM) para predizer classes CATH. O trabalho mostrou que para uma classificação binária, classes que não compartilham qualquer estrutura secundária podem ser classificadas com até 90% de precisão. Por outro lado, o trabalho também mostrou que esta precisão diminui para menos de 70% se as classes contiverem elementos estruturais comuns. Os autores destacaram ainda a baixa eficiência da SVM para classificação multiclasses, onde a precisão geral alcançou cerca de 52%.

Ao considerar estruturas proteicas há sempre o interesse em produzir modelos que capturem características relevantes e ao mesmo tempo permitam reduzir o tamanho e a complexidade dos dados, seja pela diminuição de ruídos ou por melhorias no desempenho computacional. Neste contexto, a técnica CSM descrita em Pires et al. [2011] é uma importante representação. Seu algoritmo gera um vetor que representa a frequência de pares de resíduos num determinado intervalo de distâncias.

Para isso, calcula-se a distância Euclidiana entre todos os pares de Cα, define-se um intervalo de distâncias que será considerado como ponto de corte e então calcula-se a frequência dos pares de resíduos para cada distância no intervalo definido. Ainda como forma de diminuir o tamanho e a complexidade do conjunto de dados, a SVD é uma técnica utilizada na etapa de pré-processamento para a redução de dimensionalidades. No referido trabalho, os autores aplicaram CSM e SVD em diferentes experimentos com dados baseados em números da *Enzyme Commission* (EC) e em dados da base SCOP. O trabalho alcançou uma precisão de 95% para um conjunto de dados com os 950 números mais povoados da EC e até 95% de precisão e revocação para o conjunto de todos os domínios da última versão SCOP. Em Pires et al. [2013], os autores exploram interações receptor-ligante utilizando a CSM em nível atômico (aCSM), onde o vetor de distâncias gerado para cada proteína é baseado na frequência de pares de átomos e não pares de resíduos. Em ambos os trabalhos foram usados métodos de aprendizagem supervisionada.

Em Maghawry et al. [2014], os autores propuseram uma estratégia de classificação por meio de funções proteicas. Eles utilizaram a funcionalidade da proteína baseada em atividades enzimáticas através de seis superfamílias específicas e aplicaram três métodos supervisionados de classificação para a previsão da superfamília enzimática: *Naive Bayes*, *k-nearest Neighbors*, and *Random Forest*. Os resultados alcançaram uma precisão de 98% e apresentaram uma melhora de cerca de 10% em relação ao método baseado apenas em padrões de distância.

O agrupamento de dados foi utilizado em Teletin et al. [2018] como uma estratégia de classificação não supervisionada para avaliar transições internas de proteínas, com o objetivo de obter uma melhor compreensão da dinâmica proteica e sua evolução. Nesse caso, K-*means* e o agrupamento hierárquico aglomerativo foram utilizados como métodos não supervisionados.

De acordo com a revisão, não foram encontrados na literatura trabalhos semelhantes a este. Esta estratégia envolve o uso de métodos não supervisionados para a classificação estrutural de proteínas presentes em um grande conjunto de dados. Utilizou-se aqui 150.000 estruturas para a classificação no nível classe e mais de 400.000 estruturas para a classificação no nível superfamília. Combinou-se o uso da matriz de distâncias com diferentes métodos de agrupamento e comparou-se os resultados com rótulos de classes já conhecidos da hierarquia CATH.

Capítulo 3

MATERIAIS E MÉTODOS

Para classificar estruturas de proteínas utilizando técnicas de aprendizado não supervisionado, foi executado um conjunto de tarefas que envolveram desde a coleta de dados até a avaliação de grupos formados por diferentes métodos de agrupamento. A estratégia foi dividida em cinco etapas principais - seleção dos dados, modelagem, pré-processamento, análise de agrupamento e avaliação.

As seções seguintes apresentam de forma detalhada cada uma dessas etapas e como suas tarefas foram executadas para concretizar o modelo de classificação por meio de aprendizado não supervisionado.

3.1 Seleção dos dados

A etapa de seleção é responsável por fornecer os dados usados no processo de aprendizagem. Inicialmente, foram coletadas no *Protein Data Bank* (PDB[1]) mais de 150 mil estruturas. Posteriormente, foram selecionados na base de classificação estrutural CATH[2] os descritores de classificação de 434.857 domínios, distribuídos em 4 classes, 41 arquiteturas, 1.391 topologias e 6.119 superfamílias homólogas. Tais descritores são aplicados na etapa de avaliação para validar a equivalência dos grupos formados com as classes CATH. A seleção dos dados (PDB e CATH) ocorreu em julho de 2018. Por fim, são construídas nesta etapa duas bases de dados - uma contendo as informações estruturais coletadas no PDB (Figura 3.1-a) e outra com os descritores de classificação (Figura 3.1-b).

[1]https://www.rcsb.org/
[2]http://www.cathdb.info/

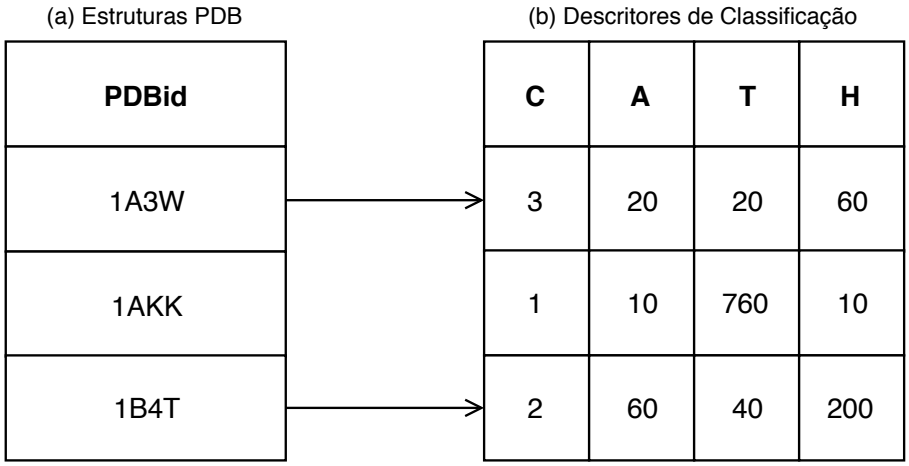


Figura 3.1. Representação das bases de dados produzidas na etapa de seleção.

3.2 Modelagem

Nesta etapa, a base produzida anteriormente com os dados coletados no PDB (Figura 3.1-a) foi submetida ao algoritmo da CSM, onde gerou-se um vetor de distâncias para cada estrutura coletada. Uma faixa com distâncias variando de 0.0 Å a 30.0 Å em intervalos de 0.2 Å foi definida, conforme Pires et al. [2011], obtendo assim uma matriz com 151 colunas que representam a quantidade de contatos em cada distância definida (Figura 3.2).

3.3 Pré-processamento

Aqui o objetivo é preparar a matriz produzida anteriormente para o processo de aprendizagem, visando melhor qualidade dos grupos formados e maior eficiência durante a execução dos métodos de agrupamento.

Inicialmente foram eliminadas as duas primeiras colunas, contendo apenas valores nulos ("0") - considerando a base de dados inicial, economizou-se o processamento de 869.714 células (2 colunas x 434.857 linhas). Posteriormente, os dados foram normalizados e submetidos à redução de dimensionalidade com SVD. Em ambos os casos foram usadas funções disponíveis na biblioteca *Scikit-learn*[3].

Para reduzir a dimensionalidade dos dados, foi analisada a distribuição dos valores singulares gerados pela SVD e definida a quantidade mais adequada de componentes (número de dimensões). Para a base utilizada neste modelo foram definidos 5 componentes, valor capaz de representar os dados originais através de

[3]http://scikit-learn.org/stable/

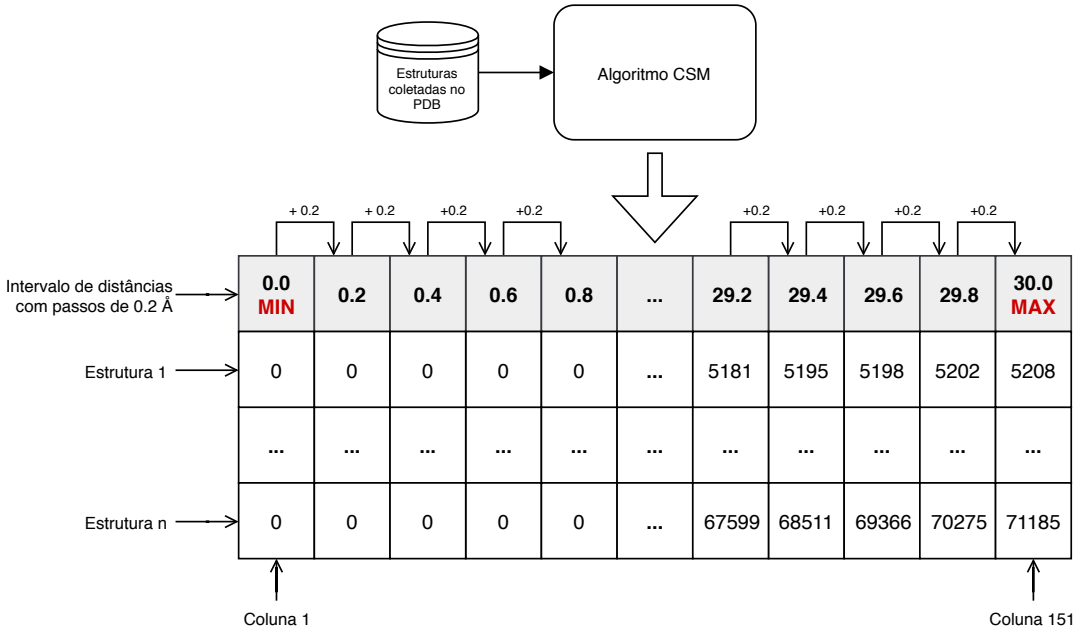


Figura 3.2. Construção da matriz de distâncias para as estruturas coletadas no PDB.

um conjunto bem menor. Deixa-se, por exemplo, de processar 65.663.407 (151 x 434.857) registros e processa-se 2.174.285 (5 x 434.857). A Figura 3.3 ilustra o pré-processamento da matriz de distâncias para a obtenção de uma nova matriz, normalizada e de dimensões reduzidas (5 componentes).

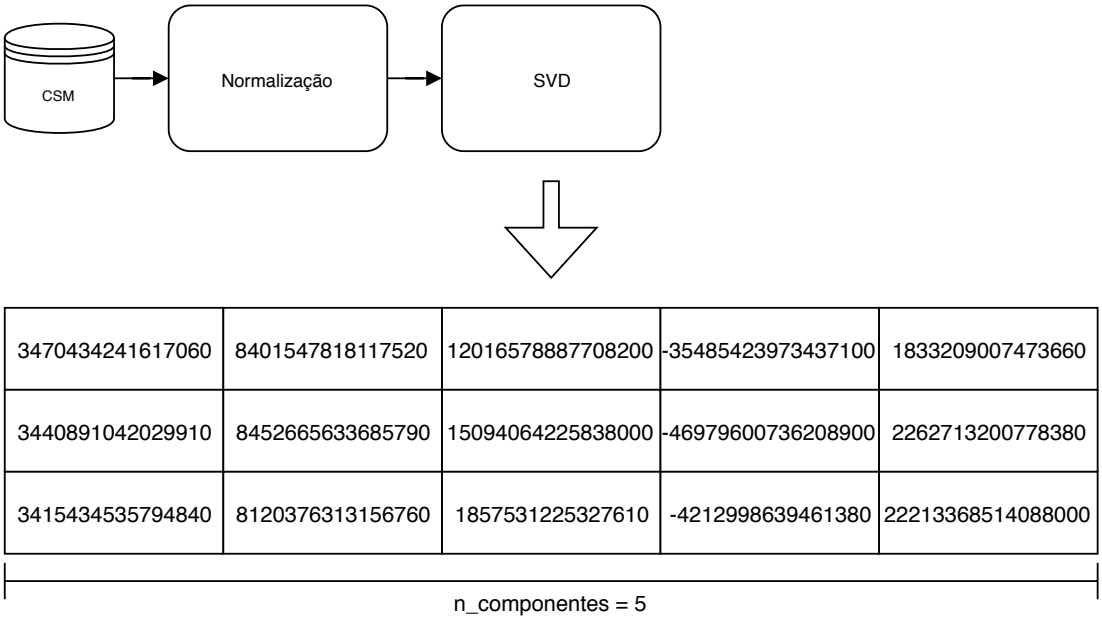


Figura 3.3. Pré-processamento da matriz de distâncias.

3.4 Análise de agrupamento e validação

Estas são as principais etapas do modelo. Aqui os métodos de agrupamento são aplicados e então avalia-se a qualidade dos grupos formados e a compatibilidade desses grupos com conjuntos de dados na hierarquia CATH. Embora sejam duas etapas distintas (agrupamento e avaliação), na prática elas são realizadas de forma conjunta, pois para cada método aplicado valida-se seus resultados.

É neste estágio que os descritores de classificação são utilizados. Eles são necessários particularmente em dois casos - para comparação dos grupos formados com conjuntos CATH; e para dividir a base de dados por topologia (C-A-T), passo importante na busca por superfamílias (H).

Esta estratégia concentra-se especificamente em tratar agrupamentos por classe (C), como uma estratégia geral de agrupamento, e superfamília (H), onde as proteínas compartilham semelhanças sequenciais, estruturais e funcionais, o que sugere a formação de bons grupos. Para isso, foram selecionados cinco métodos de agrupamento envolvendo diferentes paradigmas (Tabela 3.1).

Tabela 3.1. Métodos usados para o agrupamento dos dados

Algoritmo	Paradigma
K-*means*	Particional
Complete-Link	Hierárquico Aglomerativo
Ward	Hierárquico Aglomerativo
DBSCAN	Densidade
SpectralClustering	Espectral

Para avaliar os grupos gerados por cada algoritmo, foram aplicadas as seguintes métricas: Coeficiente de Silhueta para avaliação interna e FMI para avaliação externa. Como forma de complementar a verificação de compatibilidade com os conjuntos CATH, analisou-se ainda a homogeneidade e completude dos grupos.

3.4.1 Agrupamento por classe

Para analisar os grupos por classe, selecionou-se aleatoriamente uma amostra com 150.000 estruturas na base CSM, já normalizada e com dimensões reduzidas. A seleção de um conjunto de dados menor é um fator bastante relevante neste caso, pois alguns algoritmos de agrupamento, como DBSCAN e *SpectralClustering*, exigem um alto poder de processamento, sendo inviável trabalhar com a base completa.

Como o objetivo é separar os dados em quatro conjuntos correspondentes às classes CATH, os algoritmos foram previamente configurados para quatro grupos (n-grupos = 4). Os resultados de cada método foram então submetidos às métricas Silhueta, FMI, Homogeneidade e Completude. A Figura 3.4 apresenta as tarefas realizadas para a análise de agrupamento por classe.

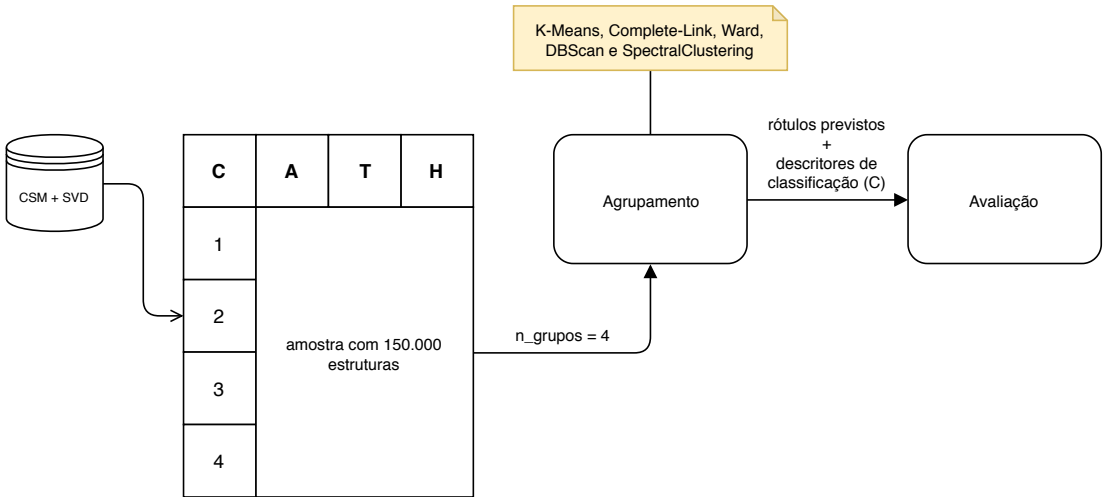


Figura 3.4. Análise de agrupamento por classe (C).

Especificamente para o método DBSCAN, foram realizados diferentes experimentos com a propriedade EPS variando entre 1 e 10, e para cada valor de EPS, calculou-se o Coeficiente de Silhueta.

3.4.2 Agrupamento por superfamília

O agrupamento por superfamília é uma tarefa mais complexa se comparada ao agrupamento por classe, já que não limita-se a uma única execução por algoritmo e não há um valor fixo para o número de grupos a serem formados. Como a base CATH possui 1.391 topologias, seriam necessárias 6.955 execuções (1.391 x 5 algoritmos). Dessa forma, para cada execução haveria uma quantidade diferente de grupos, de acordo com o número de superfamílias em cada topologia.

No entanto, para este caso há uma condição: o agrupamento deve ser implementado apenas em topologias contendo mais de uma família, ou seja, mais de um grupo - esta é uma exigência dos algoritmos de agrupamento. Assim, o primeiro passo foi selecionar as topologias que atendem a esse critério, sendo criada uma nova tabela com os descritores de classificação C-A-T-H gerados na etapa de seleção. Posteriormente um script SQL foi definido e executado para encontrar as

estruturas desejadas (Figura 3.5), o que retornou, junto aos descritores, 3.926 superfamílias e 365.996 exemplares distribuídos em 422 topologias. Esses dados foram utilizados como entrada no processo de agrupamento.

```
SELECT c,a,t, COUNT(*) AS n_H, SUM(samples) AS n_samples FROM
(
    SELECT c,a,t,h, COUNT(*) AS samples FROM db_cath.mytable
    GROUP BY c,a,t,h
) AS temp

GROUP BY c,a,t
HAVING n_H >= 2
```

Figura 3.5. Script SQL utilizado na seleção de topologias.

Para cada uma das topologias retornadas, a base CSM foi filtrada de acordo com os descritores C-A-T e em cada execução foram selecionados apenas os exemplares daquela topologia. Os dados foram então submetidos aos métodos de agrupamento com o número de grupos configurado conforme a quantidade de superfamílias. Por fim, foram aplicadas as métricas de avaliação para validar os resultados de cada algoritmo. Ao todo foram realizadas 1.688 execuções (422 x 4 algoritmos - aqui, descartou-se o método DBSCAN devido aos resultados no agrupamento por classe). A Figura 3.6 apresenta as tarefas realizadas no agrupamento por superfamília.

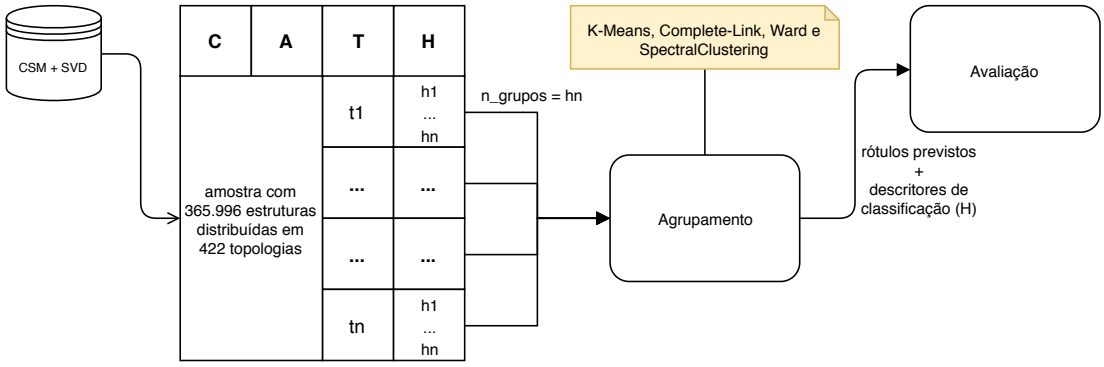


Figura 3.6. Análise de agrupamento por superfamília (H).

Capítulo 4

RESULTADOS E DISCUSSÕES

Conforme descrito em 3.4, para avaliar o desempenho dos métodos de agrupamento no nível classe (C) e também no nível superfamília (H), foram extraídos o Coeficiente de Silhueta para avaliação interna e o índice *Fowlkes Mallows*, a homogeneidade e completude para avaliação externa dos grupos.

4.1 Agrupamento por classe

A Tabela 4.1 apresenta os resultados do agrupamento por classe com cada um dos métodos aplicados. Eles foram obtidos distribuindo as estruturas presentes na CSM em quatro grupos, e comparando esses grupos com as classes do primeiro nível na hierarquia CATH (veja 3.4.1).

Tabela 4.1. Agrupamento por classe: resultados com diferentes métodos.

Método	Silhueta (-1 a 1)	FMI (0 a 1)	Homogeneidade (0 a 1)	Completude (0 a 1)	V-*measure* (0 a 1)
K-*means*	0.6031	0.3805	0.0003	0.0003	0.0003
Complete-Link	0.8439	0.5801	0.0001	0.0196	0.0003
Ward	0.6664	0.4387	0.0003	0.0005	0.0004
DBSCAN	-0.2905	0.5857	0.0002	0.0181	0.0004
SpectralClustering	-0.4547	0.5853	0.0003	0.0167	0.0006

Em relação à qualidade dos grupos formados, medida por meio do Coeficiente de Silhueta, os resultados mostram que apenas o método *Complete-Link* realizou uma boa alocação dos objetos, sendo o que mais se aproximou de 1 (0.8439). Por outro lado, os resultados alcançados com os métodos DBSCAN e *SpectralClustering* foram baixos, embora estes dois métodos tenham obtido um FMI superior aos de-

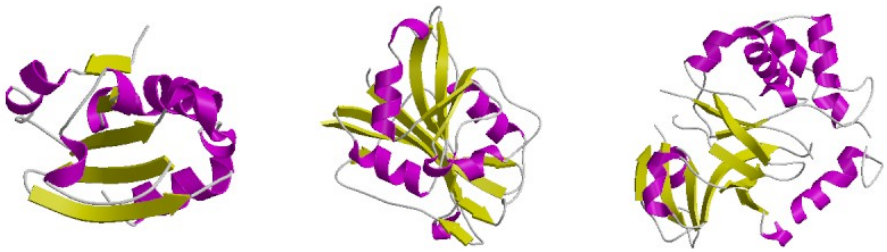
mais. Tal ocorrência sugere a aplicação de outras métricas internas para avaliá-los. Este trabalho limitou-se à aplicação do Coeficiente de Silhueta.

Quanto à comparação dos rótulos obtidos, os resultados indicam uma baixa compatibilidade entre os grupos formados e as quatro classes CATH. Ainda se considerados os melhores valores gerados pelo FMI, nenhum método obteve 60% ou mais de equivalência - DBSCAN e *SpectralClustering* (0.585) e *Complete-Link* (0.580). Isto sugere atribuições de rótulo bastante independentes. Ademais, os valores alcançados para a homogeneidade e completude ficaram bem próximos de zero em todos os métodos, o que significa muitos grupos formados por estruturas de diferentes classes (grupos heterogêneos), e que nem todas as estruturas de uma classe foram atribuídas ao mesmo grupo.

Como exemplo, a Tabela 4.2 apresenta três domínios que compartilham a mesma classe CATH, mas foram alocados em grupos distintos por alguns métodos. Neste caso, somente *Complete-Link*, DBSCAN e *SpectralClustering* preservaram as estruturas em um mesmo grupo. Em seguida, a Figura 4.1 mostra a conformação dos domínios apresentados.

Tabela 4.2. Exemplo de domínios que compartilham a mesma classe CATH, mas foram alocados em grupos distintos.

Domínio	Classe		Rótulo Atribuído				
			K-*means*	*Complete-Link*	*Ward*	DBSCAN	*SpectralClustering*
2fchA00	3	*Alpha Beta*	3	0	0	-1	0
2w3mB00	3	*Alpha Beta*	0	0	2	-1	0
2wfdA00	3	*Alpha Beta*	2	0	2	-1	0



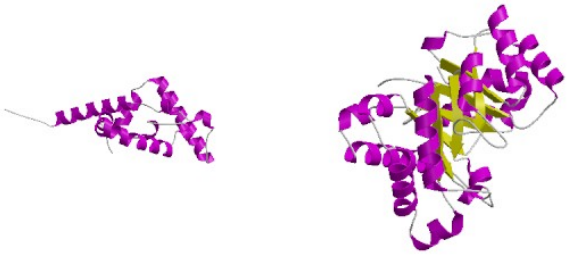
(a) (b) (c)

Figura 4.1. Exemplo de domínios que compartilham a mesma classe CATH, mas foram alocados em grupos distintos: (a) Domínio: 2fchA00, Estrutura: 2FCH - *Thioredoxin* 1. (b) Domínio: 2w3mB00, Estrutura: 2W3M - *Dihydrofolate reductase*. (c) Domínio: 2wfdA00, Estrutura: 2WFD - *Leucine*–tRNA *ligase*, *cytoplasmic*.

Por outro lado, a Tabela 4.3 mostra que os mesmos métodos - *Complete-Link*, DBSCAN e *SpectralClustering*, também alocaram em um único grupo estruturas que pertencem a classes CATH distintas. Neste caso, somente os métodos K-*means* e *Ward* deixaram as estruturas separadas. Isto reforça a hipótese de que o agrupamento por classe produz atribuições de rótulo aleatórias. A Figura 4.2 mostra a conformação dos domínios apresentados.

Tabela 4.3. Exemplo de domínios que pertencem a classes CATH distintas, mas foram alocados em um mesmo grupo.

Domínio	Classe		Rótulo Atribuído				
			K-*means*	*Complete-Link*	*Ward*	DBSCAN	*SpectralClustering*
3cx5G00	1	*Mainly Alpha*	3	0	0	-1	0
3uwvB00	3	*Alpha Beta*	2	0	2	-1	0



(a) (b)

Figura 4.2. Exemplo de domínios que pertencem a classes CATH distintas, mas foram alocados em um mesmo grupo: (a) Domínio: 3cx5G00, Estrutura: 3CX5 - *Cytochrome* b-c1 *complex subunit* 7. (b) Domínio: 3uwvB00, Estrutura: 3UWV - *Triosephosphate isomerase*.

Os resultados são complementados por meio da composição dos grupos formados em relação às classes CATH (Tabela 4.4) e da distribuição de cada classe nesses grupos (Tabela 4.5). Os dados demonstram a diversidade de estruturas incluídas em cada grupo e uma atribuição de grupo incompleta. Apenas o método *Complete-Link*, por exemplo, conseguiu gerar grupos nos quais todas as estruturas pertencem somente a uma classe (grupos B e D).

Acreditamos que a discrepância encontrada entre os grupos formados e as classes existentes, pode estar associada às características dos dados na CSM em relação à classificação CATH, onde estruturas semelhantes podem ser encontradas em classes distintas e estruturas distintas também podem ser encontradas em uma mesma classe (Tabela 4.6). Os termos "distintas"e "semelhantes"aplicados aqui, baseiam-se na frequência de contatos em cada limiar de distância.

Tabela 4.4. Composição dos grupos formados em relação às classes CATH usando diferentes métodos (homogeneidade).

Método	Classe	Grupos Formados			
		A	B	C	D
K-*means*	1	20.28%	30.50%	13.17%	32.07%
	2	24.91%	15.07%	11.66%	28.83%
	3	54.81%	54.41%	75.17%	34.99%
	4	0.00%	0.02%	0.00%	4.11%
		100.00%	100.00%	100.00%	100.00%
Complete-Link	1	26.73%	0.00%	25.00%	0.00%
	2	25.20%	0.00%	13.39%	0.00%
	3	45.62%	100.00%	61.61%	100.00%
	4	2.45%	0.00%	0.00%	0.00%
		100.00%	100.00%	100.00%	100.00%
Ward	1	28.72%	28.69%	28.70%	29.65%
	2	26.59%	27.16%	26.65%	26.03%
	3	42.91%	42.53%	42.85%	42.31%
	4	1.77%	1.62%	1.79%	2.02%
		100.00%	100.00%	100.00%	100.00%
DBSCAN	1	26.83%	22.22%	nulo	nulo
	2	25.39%	27.78%	nulo	nulo
	3	45.52%	44.44%	nulo	nulo
	4	2.26%	5.56%	nulo	nulo
		100.00%	100.00%	nulo	nulo
SpectralClustering	1	26.80%	32.00%	0.00%	50.00%
	2	25.63%	16.00%	50.00%	0.00%
	3	45.08%	48.00%	50.00%	0.00%
	4	2.49%	4.00%	0.00%	50.00%
		100.00%	100.00%	100.00%	100.00%

4.2 Agrupamento por superfamília

A Tabela 4.7 apresenta os resultados do agrupamento por superfamília. Eles foram obtidos dividindo as estruturas presentes na CSM por topologia e agrupando tais estruturas de acordo com o número de superfamílias em cada uma; os grupos formados foram então comparados com os do último nível na hierarquia CATH, utilizando os descritores de classificação (veja 3.4.2). Aqui, descartou-se o método DBSCAN devido aos resultados obtidos no agrupamento por classe - o método mostrou-se inadequado por não separar as estruturas em um número desejado de grupos. Da mesma forma, ao ser executado o método *SpectralClustering*, também

Tabela 4.5. Distribuição das classes nos grupos formados usando diferentes métodos (completude).

Método	Classe	Grupos Formados				
		A	B	C	D	
K-*means*	1	18.26%	3.96%	6.40%	71.39%	100.00%
	2	23.80%	2.08%	6.01%	68.11%	100.00%
	3	28.89%	4.13%	21.38%	45.60%	100.00%
	4	0.00%	0.03%	0.00%	99.97%	100.00%
Complete-Link	1	99.86%	0.00%	0.14%	0.00%	100.00%
	2	99.92%	0.00%	0.08%	0.00%	100.00%
	3	99.79%	0.01%	0.20%	0.00%	100.00%
	4	100.00%	0.00%	0.00%	0.00%	100.00%
Ward	1	72.53%	6.61%	19.88%	0.99%	100.00%
	2	72.41%	6.75%	19.91%	0.94%	100.00%
	3	72.61%	6.56%	19.88%	0.95%	100.00%
	4	72.70%	6.07%	20.14%	1.09%	100.00%
DBSCAN	1	99.85%	0.15%	nulo	nulo	100.00%
	2	99.80%	0.20%	nulo	nulo	100.00%
	3	99.82%	0.18%	nulo	nulo	100.00%
	4	99.56%	0.44%	nulo	nulo	100.00%
SpectralClustering	1	99.66%	0.30%	0.00%	0.04%	100.00%
	2	99.80%	0.16%	0.04%	0.00%	100.00%
	3	99.71%	0.27%	0.02%	0.00%	100.00%
	4	99.20%	0.40%	0.00%	0.40%	100.00%

Tabela 4.6. Exemplo de estruturas "aparentemente"semelhantes em classes distintas e estruturas "aparentemente"distintas em uma mesma classe.

Col.142	Col.143	Col.144	Col.145	Col.146	Col.147	Col.148	Col.149	Col.150	Col.151	Classe
8953	8990	9040	9086	9148	9203	9260	9308	9351	9392	3
8791	8862	8929	8998	9072	9127	9190	9249	9317	9392	3
8928	8975	9023	9070	9132	9193	9241	9287	9326	9392	1
8921	8978	9047	9095	9141	9196	9249	9299	9349	9392	3
8991	9043	9084	9127	9165	9209	9263	9308	9347	9392	2
1335	1341	1345	1348	1352	1356	1362	1367	1371	1376	2
75201	76235	77198	78169	79199	80266	81307	82376	83390	84472	2
755	759	760	761	764	764	764	765	766	770	2
105980	107524	109047	110540	112144	113730	115294	116920	118531	120092	2
23146	23371	23595	23855	24066	24275	24509	24715	24936	25136	2

não foram obtidos resultados satisfatórios - das 422 topologias selecionadas, a execução foi concluída em apenas 254 casos. Portanto, para a consolidação dos resultados, foram considerados os seguintes métodos: K-*means*, *Complete-Link* e *Ward*.

Ao separar as estruturas por superfamílias, foram obtidos resultados bem sucedidos se comparados ao agrupamento por classe. Considerando a média geral dos agrupamentos realizados para as 4 classes do primeiro nível, o melhor resultado foi alcançado com o método *Complete-Link*, que conseguiu produzir um bom agrupamento (Silhueta = 0.7473) e ainda obter um índice de compatibilidade igual a 75%. Se considerada apenas a primeira classe, por exemplo, um resultado ainda melhor pode ser obtido (FMI = 0.7770 e Silhueta = 0.7631).

Tabela 4.7. Agrupamento por superfamília: resultados gerais com diferentes métodos (média por classe e geral).

Método	Classe	Silhueta (-1 a 1)	FMI (0 a 1)	Homogeneidade (0 a 1)	Completude (0 a 1)	V-*measure* (0 a 1)
K-*Means*	1	0.7669	0.7452	0.5858	0.4822	0.5086
	2	0.7358	0.6877	0.5287	0.3997	0.4368
	3	0.7428	0.7283	0.5230	0.4308	0.4501
	4	0.7309	0.6827	0.2137	0.2015	0.1952
	GERAL	**0.7492**	**0.7246**	0.5402	0.4377	0.4627
Complete-Link	1	0.7631	0.7770	0.5528	0.5192	0.5152
	2	0.7454	0.7295	0.4919	0.4114	0.4302
	3	0.7375	0.7595	0.4933	0.4521	0.4517
	4	0.7081	0.6745	0.1308	0.1455	0.1178
	GERAL	**0.7473**	**0.7577**	0.5070	0.4610	0.4630
Ward	1	0.7646	0.7418	0.5784	0.4750	0.5012
	2	0.7310	0.6899	0.5334	0.4025	0.4398
	3	0.7393	0.7276	0.5207	0.4316	0.4497
	4	0.7257	0.6820	0.2115	0.2013	0.1952
	GERAL	**0.7458**	**0.7236**	0.5377	0.4363	0.4607

Foi observado que, em grande parte dos casos, o desempenho dos métodos de agrupamento esteve diretamente relacionado à quantidade de grupos a serem formados - à medida que aumenta-se os grupos, obtém-se uma queda no desempenho (Tabela 4.8). Considerando apenas as topologias com até 25 superfamílias, obteve-se resultados ainda melhores - foi possível alcançar uma média de 82% de compatibilidade aplicando o método *Complete-Link*, e 85% com o mesmo método, se considerada somente a primeira classe. Esses resultados podem ser considerados relevantes em nossa análise, já que as topologias com até 25 superfamílias representam 94% das 422 topologias selecionadas.

Por fim, para reforçar a capacidade do modelo em classificar as estruturas CATH por superfamílias usando métodos não supervisionados, a Tabela 4.9 apresenta o percentual de topologias onde a média geral para o FMI foi igual ou superior a 85%. Para este caso, os resultados alcançados são bastante interessantes - com

o método *Complete-Link*, por exemplo, em cerca de 40% das topologias houve um índice de compatibilidade igual ou superior a 85%, e quase 20% tiveram as superfamílias classificadas com 100% de compatibilidade.

Tabela 4.8. Agrupamento por superfamília: resultados com diferentes métodos para quantidades específicas de grupos (média por classe e geral).

Método	Classe	até 5 grupos		até 15 grupos		até 25 grupos	
		Silhueta (-1 a 1)	FMI (0 a 1)	Silhueta (-1 a 1)	FMI (0 a 1)	Silhueta (-1 a 1)	FMI (0 a 1)
K-*means*	1	0.8232	0.8360	0.7969	0.7899	0.7871	0.7760
	2	0.7925	0.7872	0.7699	0.7441	0.7616	0.7328
	3	0.7821	0.8021	0.7616	0.7619	0.7542	0.7464
	4	0.7309	0.6827	-	-	-	-
	GERAL	**0.7966**	**0.8078**	**0.7744**	**0.7661**	**0.7662**	**0.7523**
Complete-Link	1	0.8218	0.8555	0.7936	0.8237	0.7841	0.8122
	2	0.8042	0.8044	0.7806	0.7810	0.7726	0.7710
	3	0.7778	0.8258	0.7572	0.7902	0.7490	0.7763
	4	0.7081	0.6745	-	-	-	-
	GERAL	**0.7960**	**0.8280**	**0.7732**	**0.7973**	**0.7648**	**0.7854**
Ward	1	0.8219	0.8307	0.7953	0.7856	0.78560	0.7718
	2	0.7930	0.7891	0.7671	0.7454	0.7586	0.7343
	3	0.7805	0.8016	0.7593	0.7618	0.7521	0.7464
	4	0.7257	0.6820	-	-	-	-
	GERAL	**0.7954**	**0.8061**	**0.7722**	**0.7649**	**0.7641**	**0.7512**

Tabela 4.9. Agrupamento por superfamília: percentual de topologias com o FMI a partir de 0.85.

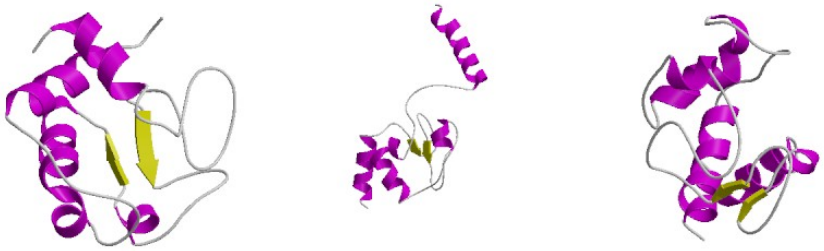
Método	0.85 a 0.89	0.9 a 0.94	0.95 a 0.99	Exatamente 1	Total
K-*means*	4.50%	7.35%	3.08%	18.48%	33.41%
Complete-Link	4.27%	9.00%	7.11%	18.96%	39.34%
Ward	4.27%	7.11%	3.79%	18.25%	33.41%

As Tabelas 4.10 e 4.11 apresentam, respectivamente, domínios que compartilham a mesma superfamília CATH e foram mantidos em um mesmo grupo, e domínios que, embora compartilhem a mesma superfamília, foram alocados em grupos distintos. As Figuras 4.3 e 4.4, mostram a conformação dos domínios apresentados.

Os resultados detalhados do agrupamento por superfamília podem ser visualizados no Apêndice A, por meio da Tabela A.1.

Tabela 4.10. Exemplo de domínios que compartilham a mesma superfamília CATH e também foram alocados em um mesmo grupo.

Domínio	Superfamília		Rótulo Atribuído		
			K-*means*	*Complete-Link*	*Ward*
2giwA00	1.10.760.10	*Cytochrome* c	0	1	1
3nbtA00	1.10.760.10	*Cytochrome* c	0	1	1
1fi9A00	1.10.760.10	*Cytochrome* c	0	1	1

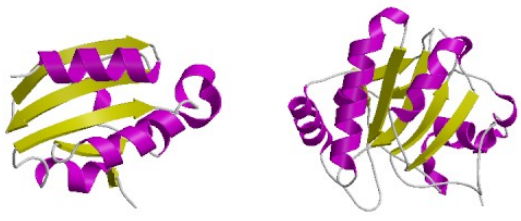


(a) (b) (c)

Figura 4.3. Exemplo de domínios que compartilham a mesma superfamília CATH e também foram alocados em um mesmo grupo. Superfamília: 1.10.760.10 - *Cytochrome* c. (a) 2giwA00 (b) 3nbtA00 (c) 1fi9A00.

Tabela 4.11. Exemplo de domínios que compartilham a mesma superfamília CATH, mas foram alocados em grupos distintos.

Domínio	Superfamília		Rótulo Atribuído		
			K-*means*	*Complete-Link*	*Ward*
2h6yA00	3.40.30.10	*Glutaredoxin*	1	0	0
2gs3A00	3.40.30.10	*Glutaredoxin*	0	0	2



(a) (b)

Figura 4.4. Exemplo de domínios que compartilham a mesma superfamília CATH, mas foram alocados em grupos distintos. Superfamília: 3.40.30.10 - *Glutaredoxin*. (a) 2h6yA00 (b) 2gs3A00.

Capítulo 5

CONCLUSÕES

A velocidade com que novas estruturas de proteínas são depositadas em bases públicas de dados, como o PDB, demanda o desenvolvimento de novas estratégias para o tratamento de grandes volumes de dados. Neste cenário, características estruturais como as disponíveis na base de dados CATH, possuem um papel fundamental na descoberta de relações estruturais e funcionais proteicas. Como forma de explorar a aplicação de técnicas não supervisionadas em dados estruturais de proteínas para a formação de grupos equivalentes à classificação CATH, foi desenvolvida uma abordagem sistemática baseada em padrões de distância inter-resíduos, envolvendo vários experimentos com diferentes técnicas de agrupamento e avaliação. A motivação para este trabalho surgiu de duas questões principais:

1. É possível gerar bons agrupamentos a partir da matriz de distâncias gerada pelo algoritmo da CSM?

2. Os grupos formados são similares à classificação CATH?

É importante destacar que esta pesquisa abordou a classificação por classe e superfamília. Os resultados mostram duas situações distintas - um conjunto de dados bastante fragmentado, se consideradas todas as classes do primeiro nível; e um conjunto bem distribuído, se consideradas as estruturas por topologia, as quais compartilham características fundamentais para o agrupamento não supervisionado. Se de um lado a estratégia mostrou-se incapaz de separar os dados por classe, por outro, revelou uma capacidade considerável em agrupá-los por superfamília. Assim, entendemos que os objetivos de pesquisa foram alcançados: a) foram coletados as estruturas no PDB e os descritores de classificação correspondentes na base CATH; b) foi construída uma base de dados com informações estruturais (CSM) associadas

aos seus descritores; c) foi implementada uma estratégia não supervisionada, e esta aplicada em agrupamentos envolvendo um grande volume de dados (150.000 estruturas no nível classe e 434.857 estruturas no nível superfamília); e os grupos foram avaliados em relação à classificação CATH.

5.1 Trabalhos Futuros

Uma avaliação mais profunda do uso de técnicas não supervisionadas para a classificação estrutural de proteínas requer a experimentação em outros níveis dentro da hierarquia CATH e em outras bases de classificação estrutural. Assim, como parte de estudos futuros, pretendemos explorar os demais níveis de classificação da base CATH (A-T) e validar a estratégia por meio de outras bases, como a *Structural Classification of Proteins* (SCOP). Pretendemos ainda, desenvolver uma estratégia de enriquecimento dos dados da CSM, de maneira que ela forneça padrões ainda mais significativos, e também uma estratégia de visualização de dados específica para a representação desses padrões.

Referências Bibliográficas

Bach, F. R. & Jordan, M. I. (2003). Learning spectral clustering. In *Advances in Neural Information Processing Systems 16*. MIT Press.

Buxbaum, E. (2015). *Fundamental of Protein Structure and Function*. Springer.

Cios, K. J.; Pedrycz, W.; Swiniarski, R. W. & Kurgan, L. A. (2007). *A Knowledge Discovery Approach*. Springer.

Cuff, A.; Sillitoe, I.; Lewis, T. E.; Redfern, O.; Garratt, R.; Thornton, J. M. & Orengo, C. A. (2009). The cath classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Research*, 37:D310–D314.

de Castro, L. N. & Ferrari, D. G. (2016). *Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações*. Saraiva, São Paulo.

Dziopa, T. (2016). Clustering validity indices evaluation with regards to semantic homogeneity. *FedCSIS*, 9:3–9.

Elden, L. (2007). *Matrix Methods in Data Mining and Pattern Recognition*. Fundamentals of Algorithms. SIAM, Philadelphia.

Fayyad, U.; Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54.

Fisher, D. H.; Pazzani, M. J. & Langley, P. (2014). *Concept Formation: Knowledge and Experience in Unsupervised Learning*. Morgan Kaufman Publishers.

Fromm, H. J. & Hargrove, M. (2012). *Essentials of biochemistry*. Springer.

Gao, X.; Chen, J. Y. & Zaki, M. J. (2018). Multiscale and multimodal analysis for computational biology. *IEEE/ACM Transactions on Computational Biology and Bioinformatic*, 15:1951–1952.

Greene, L. H. & Lewis, T. E. (2007). The cath domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Research*, 35:D291–D297.

Gu, J. & Bourne, P. E. (2011). *Structural bioinformatics*. Wiley-Blackwell, 2ed edição.

Han, J. & Kamber, M. (2006). *Data Mining: Concepts and Techniques*. Elsevier, 2 edição.

Jain, A. K. & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, New Jersey.

Knudsen, M. & Wiuf, C. (2010). The cath database. *Human Genomics*, 4(3):207.

Lee, D.; Redfern, O. & Orengo, C. A. (2007). Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology*, 8:995–1005.

Maghawry, H. A.; Mostafa, M. G. & Gharib, T. F. (2014). A new protein structure representation for efficient protein function prediction. *J Comput Biol*.

Marzzoco, A. & Torres, B. B. (2015). *Bioquímica Básica*. Guanabara Koogan, Rio de Janeiro, 4 edição.

Mishra, B. K. & Rath, A. K. (2018). Improving the efficacy of clustering by using far enhanced clustering algorithm. *International Journal of Data Mining, Modelling and Management*, 10.

Nelson, D. & Cox, M. (2014). *Princípios de Bioquímica de Lehninger*. Artmed, Porto Alegre, 6ed edição.

Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B. & Thornton, J. M. (1997). Cath – a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109.

Pacheco, E. R. (2015). *Unsupervised Learning whit R*. Packt Publishing.

Pearl, F.; Bennett, C. F.; Bray, J. E.; Harrison, A. P.; Martin, N.; Shepherd, A. J.; Sillitoe, I.; Thornton, J. M. & Orengo, C. A. (2003). The cath database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Research*, 31(1):452–455.

Pearl, F.; Martin, N.; Bray, J. E.; Buchan, D. W. A.; Harrison, A. P.; Lee, D.; Reeves, G. A.; Shepherd, A. J.; Sillitoe, I.; Todd, A. E.; Thornton, J. M. & Orengo, C. A. (2001). A rapid classification protocol for the cath domain database to support structural genomics. *Nucleic Acids Research*, 29(1):223–227.

Pires, D. E. V.; de Melo-Minardi, R. C.; da C H da Silveira; Campos, F. F. & Meira Jr., W. (2013). acsm: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics*, 29(7):855–861.

Pires, D. E. V.; de Melo-Minardi, R. C.; dos Santos, M. A.; da Silveira, C. H.; Santoro, M. M. & Meira Jr., W. (2011). Cutoff scanning matrix (csm): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics*, 12(4):S12.

Rogen, P. & Fain, B. (2003). Automatic classification of protein structure by using gauss integrals. *Proc Natl Acad Sci U S A*.

Rosenberg, A. & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 410–420. Association for Computational Linguistics.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53--65.

scikit learn (2019). scikit-learn: Machine Learning in Python. Disponível em: `https://scikit-learn.org/stable/`. Accessado em: 10 de mai. 2019.

Sillitoe, I.; Dawson, N.; Thornton, J. M. & Orengo, C. A. (2015). The history of the cath structural classification of protein domains. *Biochimie*, 119:209–217.

Sun, X. D. & Huang, R. B. (2006). Prediction of protein structural classes using support vector machines. *Amino Acids*.

Swindells, M. B.; Orengo, C. A.; Jones, D. T.; Hutchinson, E. G. & Thornton, J. M. (1998). Contemporary approaches to protein structure classification. *BioEssays*, 20:884–891.

Tan, P.; Steinbach, M. & Kumar, V. (2005). *Introduction to Data Mining*. Pearson.

Teletin, M.; Czibula, G.; Albert, S. & Bocicor, M. (2018). Using unsupervised learning methods for enhancing protein structure insight. *ScienceDirect*, 126:19--128.

Thornton, G. (2018). Thornton Group. Disponível em: `https://www.ebi.ac.uk/research/thornton`. Accessado em: 06 de dez. 2018.

Trefethen, L. & III, D. D. (1997). *Numerical Linear Algebra*. SIAM, Philadelphia.

Voet, D.; Voet, J. G. & Pratt, C. W. (2014). *Fundamentos de Bioquímica: A Vida em Nível Molecular*. Artmed.

von Luxburg, U. (2007). A tutorial on spectral clustering clustering algorithm. *Stat. Comput.*, 4(17):395--416.

Williams, M. & Daviter, T. (2013). *Protein-Ligand Interactions*. Springer.

Apêndice A

Detalhes por topologia

A Tabela A.1 apresenta os resultados do agrupamento por superfamília detalhados por topologia: método utilizado, classe (C), arquitetura (A), topologia (T), quantidade de superfamílias em cada topologia (Grupos), Coeficiente de Silhueta (S), FMI (F), homogeneidade (H), completude (Co.) e média harmônica (V).

Tabela A.1. Resultados por topologia.

Método	C	A	T	Grupos	S	F	H	Co.	V
K-Means	1	10	8	115	0,4830	0,2534	0,6649	0,4797	0,5574
Complete	1	10	8	115	0,4705	0,3441	0,5533	0,5061	0,5286
Ward	1	10	8	115	0,4506	0,2692	0,6590	0,4881	0,5608
K-Means	1	10	10	221	0,3472	0,1230	0,5620	0,3067	0,3968
Complete	1	10	10	221	0,3784	0,2118	0,4541	0,3394	0,3884
Ward	1	10	10	221	0,3224	0,1234	0,5596	0,3073	0,3968
K-Means	1	10	12	9	0,6172	0,5290	0,6796	0,2345	0,3487
Complete	1	10	12	9	0,4354	0,5948	0,6665	0,2478	0,3613
Ward	1	10	12	9	0,6109	0,5270	0,6760	0,2324	0,3459
K-Means	1	10	15	2	0,9983	1,0000	1,0000	1,0000	1,0000
Complete	1	10	15	2	0,9983	1,0000	1,0000	1,0000	1,0000
Ward	1	10	15	2	0,9983	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	20	14	0,6490	0,5093	0,5820	0,2306	0,3303
Complete	1	10	20	14	0,7323	0,8674	0,3859	0,4948	0,4336
Ward	1	10	20	14	0,6399	0,4986	0,5649	0,2218	0,3185
K-Means	1	10	30	4	0,8782	0,9675	0,9730	0,3803	0,5469
Complete	1	10	30	4	0,8782	0,9675	0,9730	0,3803	0,5469
Ward	1	10	30	4	0,8782	0,9675	0,9730	0,3803	0,5469
K-Means	1	10	40	8	0,7124	0,6482	0,9079	0,4806	0,6285
Complete	1	10	40	8	0,5143	0,8883	0,9360	0,6774	0,7860
Ward	1	10	40	8	0,7030	0,6610	0,8871	0,4802	0,6231
K-Means	1	10	60	4	0,8918	0,8102	0,6148	0,6602	0,6367

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Complete	1	10	60	4	0,8918	0,8102	0,6148	0,6602	0,6367
Ward	1	10	60	4	0,8918	0,8102	0,6148	0,6602	0,6367
K-Means	1	10	100	3	0,5756	0,6802	0,3661	0,1428	0,2054
Complete	1	10	100	3	0,5764	0,6813	0,3678	0,1435	0,2065
Ward	1	10	100	3	0,5764	0,6813	0,3678	0,1435	0,2065
K-Means	1	10	110	3	0,9591	1,0000	1,0000	1,0000	1,0000
Complete	1	10	110	3	0,9591	1,0000	1,0000	1,0000	1,0000
Ward	1	10	110	3	0,9591	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	132	13	0,7921	0,5896	0,8014	0,6704	0,7300
Complete	1	10	132	13	0,7409	0,6349	0,7577	0,7104	0,7333
Ward	1	10	132	13	0,7921	0,5896	0,8014	0,6704	0,7300
K-Means	1	10	150	67	0,4612	0,1806	0,5521	0,3940	0,4598
Complete	1	10	150	67	0,4618	0,2190	0,4843	0,4141	0,4465
Ward	1	10	150	67	0,4465	0,1803	0,5543	0,3916	0,4589
K-Means	1	10	167	3	0,9375	0,9212	0,8962	0,1782	0,2973
Complete	1	10	167	3	0,9375	0,9212	0,8962	0,1782	0,2973
Ward	1	10	167	3	0,9375	0,9212	0,8962	0,1782	0,2973
K-Means	1	10	168	2	0,8538	0,7780	0,0033	0,0017	0,0022
Complete	1	10	168	2	0,8538	0,7780	0,0033	0,0017	0,0022
Ward	1	10	168	2	0,8538	0,7780	0,0033	0,0017	0,0022
K-Means	1	10	210	2	0,9773	1,0000	1,0000	1,0000	1,0000
Complete	1	10	210	2	0,9773	1,0000	1,0000	1,0000	1,0000
Ward	1	10	210	2	0,9773	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	220	12	0,6302	0,6053	0,7176	0,5507	0,6232
Complete	1	10	220	12	0,6519	0,5944	0,6240	0,5258	0,5707
Ward	1	10	220	12	0,6838	0,5917	0,6681	0,5469	0,6014
K-Means	1	10	225	4	0,7211	0,6484	0,0849	0,0293	0,0436
Complete	1	10	225	4	0,5616	0,6517	0,3320	0,0989	0,1525
Ward	1	10	225	4	0,7157	0,6557	0,0820	0,0289	0,0428
K-Means	1	10	230	2	0,6049	0,6915	0,1246	0,0165	0,0292
Complete	1	10	230	2	0,6419	0,7944	0,0480	0,0085	0,0145
Ward	1	10	230	2	0,6049	0,6915	0,1246	0,0165	0,0292
K-Means	1	10	238	16	0,5611	0,4745	0,4010	0,0758	0,1274
Complete	1	10	238	16	0,6685	0,7824	0,3017	0,1244	0,1761
Ward	1	10	238	16	0,5796	0,5306	0,4165	0,0870	0,1440
K-Means	1	10	246	20	0,6379	0,5928	0,7509	0,7124	0,7312
Complete	1	10	246	20	0,4600	0,4767	0,6870	0,6876	0,6873
Ward	1	10	246	20	0,6129	0,5690	0,7244	0,7017	0,7129
K-Means	1	10	260	16	0,5610	0,4714	0,4216	0,1647	0,2369
Complete	1	10	260	16	0,6440	0,5666	0,4018	0,2068	0,2730
Ward	1	10	260	16	0,5858	0,4059	0,4198	0,1527	0,2240

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
K-Means	1	10	269	2	0,8152	1,0000	1,0000	1,0000	1,0000
Complete	1	10	269	2	0,8152	1,0000	1,0000	1,0000	1,0000
Ward	1	10	269	2	0,8152	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	274	9	0,8028	0,8397	0,8056	0,8151	0,8103
Complete	1	10	274	9	0,8251	0,7418	0,6258	0,8863	0,7336
Ward	1	10	274	9	0,7953	0,8412	0,8082	0,8246	0,8163
K-Means	1	10	275	3	0,8927	0,7363	0,6808	0,2882	0,4050
Complete	1	10	275	3	0,8927	0,7363	0,6808	0,2882	0,4050
Ward	1	10	275	3	0,8927	0,7363	0,6808	0,2882	0,4050
K-Means	1	10	286	10	0,8943	0,7056	0,8241	0,7023	0,7583
Complete	1	10	286	10	0,8759	0,7320	0,8241	0,7468	0,7835
Ward	1	10	286	10	0,8943	0,7056	0,8241	0,7023	0,7583
K-Means	1	10	287	92	0,4832	0,2690	0,5841	0,5460	0,5644
Complete	1	10	287	92	0,4779	0,2393	0,4935	0,5623	0,5256
Ward	1	10	287	92	0,4509	0,2714	0,5851	0,5435	0,5635
K-Means	1	10	288	2	0,7834	0,5813	0,3275	0,2365	0,2746
Complete	1	10	288	2	0,7834	0,5813	0,3275	0,2365	0,2746
Ward	1	10	288	2	0,7834	0,5813	0,3275	0,2365	0,2746
K-Means	1	10	340	5	0,7620	0,4962	0,5978	0,2304	0,3326
Complete	1	10	340	5	0,6745	0,6558	0,8056	0,3726	0,5095
Ward	1	10	340	5	0,7524	0,5115	0,5978	0,2341	0,3364
K-Means	1	10	357	11	0,5447	0,4953	0,5820	0,2041	0,3022
Complete	1	10	357	11	0,3915	0,6684	0,5111	0,2559	0,3410
Ward	1	10	357	11	0,5018	0,5088	0,5766	0,2086	0,3064
K-Means	1	10	390	5	0,8172	0,5378	0,3767	0,1849	0,2480
Complete	1	10	390	5	0,8422	0,5353	0,3608	0,1785	0,2388
Ward	1	10	390	5	0,8350	0,5387	0,3722	0,1869	0,2488
K-Means	1	10	400	2	0,8749	1,0000	1,0000	1,0000	1,0000
Complete	1	10	400	2	0,8749	1,0000	1,0000	1,0000	1,0000
Ward	1	10	400	2	0,8749	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	405	3	0,7331	0,6534	0,3692	0,1077	0,1668
Complete	1	10	405	3	0,6074	0,9568	0,4244	0,7575	0,5440
Ward	1	10	405	3	0,7281	0,6563	0,3786	0,1113	0,1720
K-Means	1	10	418	8	0,5784	0,5302	0,3329	0,0745	0,1217
Complete	1	10	418	8	0,5799	0,7020	0,1978	0,0657	0,0986
Ward	1	10	418	8	0,6661	0,6349	0,3763	0,0971	0,1544
K-Means	1	10	437	3	0,5354	0,6444	0,2293	0,1247	0,1615
Complete	1	10	437	3	0,5349	0,6451	0,2304	0,1253	0,1623
Ward	1	10	437	3	0,5862	0,6730	0,2929	0,1388	0,1883
K-Means	1	10	439	2	0,9705	0,7966	0,0467	0,0180	0,0260
Complete	1	10	439	2	0,9705	0,7966	0,0467	0,0180	0,0260

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Ward	1	10	439	2	0,9705	0,7966	0,0467	0,0180	0,0260
K-Means	1	10	443	2	0,7874	0,6167	0,3193	0,2499	0,2804
Complete	1	10	443	2	0,7874	0,6167	0,3193	0,2499	0,2804
Ward	1	10	443	2	0,7874	0,6167	0,3193	0,2499	0,2804
K-Means	1	10	472	13	0,6732	0,5982	0,6998	0,4749	0,5658
Complete	1	10	472	13	0,7003	0,6186	0,7171	0,5187	0,6020
Ward	1	10	472	13	0,7152	0,6232	0,7557	0,5153	0,6128
K-Means	1	10	490	15	0,6032	0,6166	0,5008	0,1999	0,2857
Complete	1	10	490	15	0,6557	0,7873	0,3540	0,2901	0,3189
Ward	1	10	490	15	0,5518	0,4783	0,5005	0,1702	0,2540
K-Means	1	10	510	2	0,6212	0,7574	0,0546	0,0023	0,0044
Complete	1	10	510	2	0,5968	0,8087	0,0373	0,0018	0,0035
Ward	1	10	510	2	0,6198	0,7693	0,0499	0,0021	0,0041
K-Means	1	10	520	5	0,9142	0,9085	0,6956	0,8249	0,7547
Complete	1	10	520	5	0,9142	0,9085	0,6956	0,8249	0,7547
Ward	1	10	520	5	0,9142	0,9085	0,6956	0,8249	0,7547
K-Means	1	10	530	5	0,7518	0,8432	0,6880	0,4468	0,5417
Complete	1	10	530	5	0,7914	0,7011	0,0415	0,0889	0,0566
Ward	1	10	530	5	0,7427	0,8189	0,6456	0,4121	0,5031
K-Means	1	10	533	3	0,8865	0,9215	0,0101	0,0065	0,0080
Complete	1	10	533	3	0,9686	0,9678	0,0010	0,0040	0,0016
Ward	1	10	533	3	0,8865	0,9215	0,0101	0,0065	0,0080
K-Means	1	10	606	2	0,8734	1,0000	1,0000	1,0000	1,0000
Complete	1	10	606	2	0,8734	1,0000	1,0000	1,0000	1,0000
Ward	1	10	606	2	0,8734	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	630	2	0,6820	0,7009	0,1204	0,0085	0,0159
Complete	1	10	630	2	0,4120	0,8231	0,2770	0,0267	0,0487
Ward	1	10	630	2	0,6818	0,7009	0,1170	0,0083	0,0154
K-Means	1	10	645	2	0,9461	0,9140	0,0017	0,0141	0,0030
Complete	1	10	645	2	0,9461	0,9140	0,0017	0,0141	0,0030
Ward	1	10	645	2	0,9461	0,9140	0,0017	0,0141	0,0030
K-Means	1	10	720	10	0,6052	0,3689	0,4899	0,3949	0,4373
Complete	1	10	720	10	0,5934	0,4568	0,4286	0,4460	0,4371
Ward	1	10	720	10	0,6207	0,3674	0,4775	0,3918	0,4305
K-Means	1	10	730	2	0,8553	0,8501	0,2956	0,0247	0,0456
Complete	1	10	730	2	0,8553	0,8501	0,2956	0,0247	0,0456
Ward	1	10	730	2	0,8553	0,8501	0,2956	0,0247	0,0456
K-Means	1	10	750	2	0,9607	0,9177	0,3262	0,5853	0,4189
Complete	1	10	750	2	0,9607	0,9177	0,3262	0,5853	0,4189
Ward	1	10	750	2	0,9607	0,9177	0,3262	0,5853	0,4189
K-Means	1	10	760	2	0,8613	0,9169	0,0125	0,0007	0,0014

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Complete	1	10	760	2	0,8676	0,9687	0,0042	0,0005	0,0009
Ward	1	10	760	2	0,8591	0,9095	0,0138	0,0007	0,0014
K-Means	1	10	790	2	0,5806	0,7072	0,2510	0,0803	0,1217
Complete	1	10	790	2	0,7525	0,9464	0,1609	0,2361	0,1914
Ward	1	10	790	2	0,5803	0,7076	0,2517	0,0805	0,1220
K-Means	1	10	890	4	0,7598	0,9559	0,9246	0,9041	0,9142
Complete	1	10	890	4	0,6722	0,6595	0,4393	0,6841	0,5350
Ward	1	10	890	4	0,7598	0,9559	0,9246	0,9041	0,9142
K-Means	1	10	1040	4	0,6999	0,7170	0,6972	0,2774	0,3969
Complete	1	10	1040	4	0,5352	0,7324	0,7682	0,2995	0,4309
Ward	1	10	1040	4	0,6999	0,7170	0,6972	0,2774	0,3969
K-Means	1	10	1050	2	0,9380	0,9455	0,6599	0,4199	0,5133
Complete	1	10	1050	2	0,9272	1,0000	1,0000	1,0000	1,0000
Ward	1	10	1050	2	0,9380	0,9455	0,6599	0,4199	0,5133
K-Means	1	10	1060	2	0,9875	0,8522	0,0280	0,0028	0,0051
Complete	1	10	1060	2	0,9875	0,8522	0,0280	0,0028	0,0051
Ward	1	10	1060	2	0,9875	0,8522	0,0280	0,0028	0,0051
K-Means	1	10	1070	2	0,8470	0,8228	0,4144	0,1726	0,2437
Complete	1	10	1070	2	0,8470	0,8228	0,4144	0,1726	0,2437
Ward	1	10	1070	2	0,8470	0,8228	0,4144	0,1726	0,2437
K-Means	1	10	1130	2	0,8569	0,8260	0,2989	0,0383	0,0679
Complete	1	10	1130	2	0,8569	0,8260	0,2989	0,0383	0,0679
Ward	1	10	1130	2	0,8569	0,8260	0,2989	0,0383	0,0679
K-Means	1	10	1200	23	0,4103	0,3600	0,3612	0,2258	0,2778
Complete	1	10	1200	23	0,3989	0,5063	0,2736	0,2967	0,2847
Ward	1	10	1200	23	0,3867	0,3634	0,4152	0,2417	0,3056
K-Means	1	10	1220	15	0,6015	0,5058	0,5405	0,4599	0,4969
Complete	1	10	1220	15	0,6300	0,7236	0,4849	0,6194	0,5439
Ward	1	10	1220	15	0,5710	0,5018	0,5775	0,4464	0,5035
K-Means	1	10	1240	7	0,8574	0,8641	0,7335	0,5190	0,6079
Complete	1	10	1240	7	0,8982	0,9011	0,6461	0,6152	0,6303
Ward	1	10	1240	7	0,8617	0,8641	0,7295	0,5238	0,6098
K-Means	1	10	1270	2	0,9433	1,0000	1,0000	1,0000	1,0000
Complete	1	10	1270	2	0,9433	1,0000	1,0000	1,0000	1,0000
Ward	1	10	1270	2	0,9433	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	1370	2	0,6185	1,0000	1,0000	1,0000	1,0000
Complete	1	10	1370	2	0,6015	0,7876	0,3464	0,4920	0,4066
Ward	1	10	1370	2	0,5624	0,6121	0,3386	0,3184	0,3282
K-Means	1	10	1410	4	0,6999	0,5614	0,3204	0,4268	0,3660
Complete	1	10	1410	4	0,6487	0,5390	0,1042	0,3526	0,1608
Ward	1	10	1410	4	0,6514	0,5303	0,2843	0,3904	0,3290

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
K-Means	1	10	1440	2	0,9982	1,0000	1,0000	1,0000	1,0000
Complete	1	10	1440	2	0,9982	1,0000	1,0000	1,0000	1,0000
Ward	1	10	1440	2	0,9982	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	1470	2	0,9824	1,0000	1,0000	1,0000	1,0000
Complete	1	10	1470	2	0,9824	1,0000	1,0000	1,0000	1,0000
Ward	1	10	1470	2	0,9824	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	1510	2	0,9938	1,0000	1,0000	1,0000	1,0000
Complete	1	10	1510	2	0,9938	1,0000	1,0000	1,0000	1,0000
Ward	1	10	1510	2	0,9938	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	1520	2	0,9484	0,8676	0,0252	0,0092	0,0134
Complete	1	10	1520	2	0,9484	0,8676	0,0252	0,0092	0,0134
Ward	1	10	1520	2	0,9484	0,8676	0,0252	0,0092	0,0134
K-Means	1	10	1580	2	0,9634	1,0000	1,0000	1,0000	1,0000
Complete	1	10	1580	2	0,9634	1,0000	1,0000	1,0000	1,0000
Ward	1	10	1580	2	0,9634	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	1620	2	0,7953	0,6247	0,0111	0,0062	0,0080
Complete	1	10	1620	2	0,6971	0,9072	0,2168	0,5176	0,3056
Ward	1	10	1620	2	0,7780	0,6231	0,0006	0,0004	0,0005
K-Means	1	10	1660	7	0,7055	0,5146	0,5594	0,5436	0,5514
Complete	1	10	1660	7	0,6821	0,4045	0,4056	0,4399	0,4221
Ward	1	10	1660	7	0,6732	0,5345	0,5651	0,5599	0,5625
K-Means	1	10	1740	12	0,7189	0,3938	0,6257	0,4351	0,5133
Complete	1	10	1740	12	0,7139	0,3406	0,4402	0,3564	0,3939
Ward	1	10	1740	12	0,7203	0,3934	0,6136	0,4295	0,5053
K-Means	1	10	1760	2	0,8360	0,7176	0,0268	0,0842	0,0407
Complete	1	10	1760	2	0,8360	0,7176	0,0268	0,0842	0,0407
Ward	1	10	1760	2	0,8360	0,7176	0,0268	0,0842	0,0407
K-Means	1	10	1790	2	0,7526	0,5164	0,2936	0,1909	0,2314
Complete	1	10	1790	2	0,7526	0,5164	0,2936	0,1909	0,2314
Ward	1	10	1790	2	0,7526	0,5164	0,2936	0,1909	0,2314
K-Means	1	10	1900	5	0,5841	0,6018	0,3519	0,3042	0,3263
Complete	1	10	1900	5	0,7194	0,6789	0,1867	0,3196	0,2357
Ward	1	10	1900	5	0,5480	0,5072	0,2712	0,2344	0,2515
K-Means	1	10	2010	2	0,9867	1,0000	1,0000	1,0000	1,0000
Complete	1	10	2010	2	0,9867	1,0000	1,0000	1,0000	1,0000
Ward	1	10	2010	2	0,9867	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	3020	2	0,7014	0,8144	0,0388	0,0197	0,0261
Complete	1	10	3020	2	0,7014	0,8144	0,0388	0,0197	0,0261
Ward	1	10	3020	2	0,7014	0,8144	0,0388	0,0197	0,0261
K-Means	1	10	3060	2	0,9451	1,0000	1,0000	1,0000	1,0000
Complete	1	10	3060	2	0,9451	1,0000	1,0000	1,0000	1,0000

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Ward	1	10	3060	2	0,9451	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	3100	2	0,8709	0,5025	0,1655	0,1989	0,1807
Complete	1	10	3100	2	0,8709	0,5025	0,1655	0,1989	0,1807
Ward	1	10	3100	2	0,8709	0,5025	0,1655	0,1989	0,1807
K-Means	1	10	3110	2	0,9223	1,0000	1,0000	1,0000	1,0000
Complete	1	10	3110	2	0,9223	1,0000	1,0000	1,0000	1,0000
Ward	1	10	3110	2	0,9223	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	3130	2	0,9059	1,0000	1,0000	1,0000	1,0000
Complete	1	10	3130	2	0,9059	1,0000	1,0000	1,0000	1,0000
Ward	1	10	3130	2	0,9059	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	3200	2	0,8618	1,0000	1,0000	1,0000	1,0000
Complete	1	10	3200	2	0,8618	1,0000	1,0000	1,0000	1,0000
Ward	1	10	3200	2	0,8618	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	3210	4	0,7799	0,6547	0,6954	0,1971	0,3071
Complete	1	10	3210	4	0,7045	0,6674	0,6885	0,2038	0,3145
Ward	1	10	3210	4	0,7628	0,6584	0,6922	0,1987	0,3088
K-Means	1	10	3230	3	0,7491	0,8194	0,7858	0,7661	0,7758
Complete	1	10	3230	3	0,7491	0,8194	0,7858	0,7661	0,7758
Ward	1	10	3230	3	0,7463	0,8427	0,8058	0,7865	0,7960
K-Means	1	10	3290	2	0,7555	0,7069	0,2173	0,0488	0,0797
Complete	1	10	3290	2	0,7555	0,7069	0,2173	0,0488	0,0797
Ward	1	10	3290	2	0,7555	0,7069	0,2173	0,0488	0,0797
K-Means	1	10	3350	2	0,7023	0,8896	0,5999	0,6908	0,6421
Complete	1	10	3350	2	0,7023	0,8896	0,5999	0,6908	0,6421
Ward	1	10	3350	2	0,7023	0,8896	0,5999	0,6908	0,6421
K-Means	1	10	3380	2	0,8561	1,0000	1,0000	1,0000	1,0000
Complete	1	10	3380	2	0,8561	1,0000	1,0000	1,0000	1,0000
Ward	1	10	3380	2	0,8561	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	3390	2	0,4622	0,7410	0,0632	0,0031	0,0059
Complete	1	10	3390	2	0,7713	0,9852	0,6438	0,2065	0,3128
Ward	1	10	3390	2	0,4854	0,7719	0,0499	0,0026	0,0050
K-Means	1	10	3450	4	0,6736	0,9527	0,9215	0,8785	0,8995
Complete	1	10	3450	4	0,6704	1,0000	1,0000	1,0000	1,0000
Ward	1	10	3450	4	0,6736	0,9527	0,9215	0,8785	0,8995
K-Means	1	10	3680	2	0,9661	1,0000	1,0000	1,0000	1,0000
Complete	1	10	3680	2	0,9661	1,0000	1,0000	1,0000	1,0000
Ward	1	10	3680	2	0,9661	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	3730	3	0,7595	1,0000	1,0000	1,0000	1,0000
Complete	1	10	3730	3	0,7595	1,0000	1,0000	1,0000	1,0000
Ward	1	10	3730	3	0,7595	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	4030	2	0,8293	1,0000	1,0000	1,0000	1,0000

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Complete	1	10	4030	2	0,8293	1,0000	1,0000	1,0000	1,0000
Ward	1	10	4030	2	0,8293	1,0000	1,0000	1,0000	1,0000
K-Means	1	10	4120	2	0,6510	0,6537	0,0001	0,0001	0,0001
Complete	1	10	4120	2	0,6000	0,8695	0,1123	0,1799	0,1383
Ward	1	10	4120	2	0,6510	0,6537	0,0001	0,0001	0,0001
K-Means	1	20	5	92	0,5292	0,4482	0,5330	0,5425	0,5377
Complete	1	20	5	92	0,4937	0,3074	0,3464	0,5246	0,4173
Ward	1	20	5	92	0,5255	0,4479	0,5367	0,5428	0,5397
K-Means	1	20	20	3	0,5883	0,7201	0,2235	0,0345	0,0598
Complete	1	20	20	3	0,7682	1,0000	1,0000	1,0000	1,0000
Ward	1	20	20	3	0,5866	0,7379	0,2131	0,0344	0,0592
K-Means	1	20	50	8	0,6618	0,5712	0,5747	0,5930	0,5837
Complete	1	20	50	8	0,9054	0,7318	0,4295	0,8327	0,5667
Ward	1	20	50	8	0,6593	0,5713	0,5747	0,5934	0,5839
K-Means	1	20	58	162	0,5286	0,4943	0,7553	0,6556	0,7019
Complete	1	20	58	162	0,5148	0,4216	0,6591	0,6647	0,6619
Ward	1	20	58	162	0,5218	0,4858	0,7449	0,6579	0,6987
K-Means	1	20	59	2	0,9248	0,8961	0,6865	0,5852	0,6318
Complete	1	20	59	2	0,9248	0,8961	0,6865	0,5852	0,6318
Ward	1	20	59	2	0,9248	0,8961	0,6865	0,5852	0,6318
K-Means	1	20	80	3	0,8380	0,7493	0,3522	0,1964	0,2522
Complete	1	20	80	3	0,8133	0,7383	0,1821	0,1048	0,1330
Ward	1	20	80	3	0,8243	0,7143	0,3116	0,1676	0,2180
K-Means	1	20	81	3	0,9014	0,5416	0,2389	0,1732	0,2008
Complete	1	20	81	3	0,8997	0,5422	0,2722	0,1990	0,2299
Ward	1	20	81	3	0,8997	0,5422	0,2722	0,1990	0,2299
K-Means	1	20	85	2	0,9293	1,0000	1,0000	1,0000	1,0000
Complete	1	20	85	2	0,9293	1,0000	1,0000	1,0000	1,0000
Ward	1	20	85	2	0,9293	1,0000	1,0000	1,0000	1,0000
K-Means	1	20	89	2	0,8732	1,0000	1,0000	1,0000	1,0000
Complete	1	20	89	2	0,8733	0,9869	0,8580	0,9032	0,8800
Ward	1	20	89	2	0,8733	0,9869	0,8580	0,9032	0,8800
K-Means	1	20	91	2	0,9561	0,9556	0,6468	0,3550	0,4584
Complete	1	20	91	2	0,9378	0,9493	0,0039	0,0068	0,0050
Ward	1	20	91	2	0,9561	0,9556	0,6468	0,3550	0,4584
K-Means	1	20	120	155	0,4472	0,2747	0,6675	0,6114	0,6383
Complete	1	20	120	155	0,4009	0,2219	0,5713	0,6050	0,5877
Ward	1	20	120	155	0,4294	0,2627	0,6676	0,6149	0,6402
K-Means	1	20	140	16	0,6783	0,5306	0,7589	0,6885	0,7220
Complete	1	20	140	16	0,6627	0,5691	0,7548	0,7066	0,7299
Ward	1	20	140	16	0,6778	0,5592	0,7642	0,7088	0,7355

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
K-Means	1	20	142	2	0,8972	1,0000	1,0000	1,0000	1,0000
Complete	1	20	142	2	0,8972	1,0000	1,0000	1,0000	1,0000
Ward	1	20	142	2	0,8972	1,0000	1,0000	1,0000	1,0000
K-Means	1	20	150	3	0,9644	1,0000	1,0000	1,0000	1,0000
Complete	1	20	150	3	0,9644	1,0000	1,0000	1,0000	1,0000
Ward	1	20	150	3	0,9644	1,0000	1,0000	1,0000	1,0000
K-Means	1	20	190	5	0,7985	0,8090	0,7932	0,4413	0,5671
Complete	1	20	190	5	0,8271	0,9797	0,7857	0,8566	0,8196
Ward	1	20	190	5	0,7336	0,7659	0,7960	0,4091	0,5404
K-Means	1	20	200	2	0,9527	1,0000	1,0000	1,0000	1,0000
Complete	1	20	200	2	0,9527	1,0000	1,0000	1,0000	1,0000
Ward	1	20	200	2	0,9527	1,0000	1,0000	1,0000	1,0000
K-Means	1	20	225	3	0,9031	1,0000	1,0000	1,0000	1,0000
Complete	1	20	225	3	0,9140	0,8846	0,4703	0,8996	0,6177
Ward	1	20	225	3	0,9031	1,0000	1,0000	1,0000	1,0000
K-Means	1	20	272	4	0,8672	1,0000	1,0000	1,0000	1,0000
Complete	1	20	272	4	0,7760	0,6682	0,8333	0,8093	0,8212
Ward	1	20	272	4	0,8672	1,0000	1,0000	1,0000	1,0000
K-Means	1	20	890	6	0,9159	0,9261	0,7158	0,9132	0,8025
Complete	1	20	890	6	0,9172	0,9258	0,6987	0,9300	0,7979
Ward	1	20	890	6	0,9153	0,9259	0,7138	0,9088	0,7996
K-Means	1	20	910	2	0,9351	0,9466	0,0016	0,0081	0,0027
Complete	1	20	910	2	0,9351	0,9466	0,0016	0,0081	0,0027
Ward	1	20	910	2	0,9351	0,9466	0,0016	0,0081	0,0027
K-Means	1	20	920	4	0,8265	0,9103	0,2468	0,0541	0,0887
Complete	1	20	920	4	0,8248	0,9070	0,1104	0,0248	0,0405
Ward	1	20	920	4	0,8196	0,9058	0,2434	0,0518	0,0855
K-Means	1	20	930	5	0,8131	0,8906	0,9061	0,7260	0,8061
Complete	1	20	930	5	0,8067	0,8019	0,7924	0,6346	0,7048
Ward	1	20	930	5	0,8131	0,8906	0,9061	0,7260	0,8061
K-Means	1	20	960	5	0,6704	0,5805	0,7175	0,4478	0,5515
Complete	1	20	960	5	0,7550	0,9118	0,7175	0,7902	0,7521
Ward	1	20	960	5	0,6547	0,5928	0,7175	0,4533	0,5556
K-Means	1	20	970	5	0,6821	0,5966	0,7291	0,3401	0,4639
Complete	1	20	970	5	0,6458	0,6779	0,6917	0,3653	0,4781
Ward	1	20	970	5	0,6044	0,5651	0,7330	0,3325	0,4575
K-Means	1	20	1050	6	0,6150	0,7796	0,5569	0,2063	0,3011
Complete	1	20	1050	6	0,7225	0,9267	0,5481	0,3936	0,4581
Ward	1	20	1050	6	0,6136	0,7527	0,5680	0,1847	0,2788
K-Means	1	20	1060	2	0,9114	1,0000	1,0000	1,0000	1,0000
Complete	1	20	1060	2	0,9114	1,0000	1,0000	1,0000	1,0000

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Ward	1	20	1060	2	0,9114	1,0000	1,0000	1,0000	1,0000
K-Means	1	20	1070	2	0,6883	0,8712	0,0218	0,0012	0,0024
Complete	1	20	1070	2	0,8865	0,9881	0,0013	0,0006	0,0008
Ward	1	20	1070	2	0,8865	0,9881	0,0013	0,0006	0,0008
K-Means	1	20	1130	2	0,7324	0,6812	0,1012	0,0406	0,0579
Complete	1	20	1130	2	0,7324	0,6812	0,1012	0,0406	0,0579
Ward	1	20	1130	2	0,7324	0,6812	0,1012	0,0406	0,0579
K-Means	1	20	1160	2	0,9548	0,8140	0,0271	0,2711	0,0492
Complete	1	20	1160	2	0,9548	0,8140	0,0271	0,2711	0,0492
Ward	1	20	1160	2	0,9548	0,8140	0,0271	0,2711	0,0492
K-Means	1	20	1220	2	0,8561	0,9175	0,4702	0,1224	0,1942
Complete	1	20	1220	2	0,8957	1,0000	1,0000	1,0000	1,0000
Ward	1	20	1220	2	0,8561	0,9175	0,4702	0,1224	0,1942
K-Means	1	20	1250	9	0,5638	0,3836	0,5009	0,2441	0,3283
Complete	1	20	1250	9	0,4797	0,4714	0,4597	0,2764	0,3452
Ward	1	20	1250	9	0,5308	0,3833	0,4943	0,2419	0,3248
K-Means	1	20	1260	13	0,5197	0,4663	0,4983	0,0874	0,1488
Complete	1	20	1260	13	0,3687	0,6787	0,4404	0,1225	0,1917
Ward	1	20	1260	13	0,4875	0,4854	0,5058	0,0952	0,1603
K-Means	1	20	1270	27	0,6593	0,5540	0,7640	0,6962	0,7285
Complete	1	20	1270	27	0,6669	0,4000	0,5861	0,6659	0,6234
Ward	1	20	1270	27	0,6783	0,6089	0,7951	0,7254	0,7587
K-Means	1	20	1280	24	0,6472	0,4747	0,7618	0,6739	0,7151
Complete	1	20	1280	24	0,6763	0,5232	0,7485	0,7251	0,7366
Ward	1	20	1280	24	0,6487	0,4628	0,7717	0,6716	0,7182
K-Means	1	20	1290	3	0,7937	0,6618	0,6173	0,5610	0,5878
Complete	1	20	1290	3	0,7777	0,6705	0,6493	0,5664	0,6050
Ward	1	20	1290	3	0,7990	0,6728	0,6048	0,5688	0,5863
K-Means	1	20	1300	2	0,9048	0,9175	0,6750	0,5239	0,5899
Complete	1	20	1300	2	0,9048	0,9175	0,6750	0,5239	0,5899
Ward	1	20	1300	2	0,9048	0,9175	0,6750	0,5239	0,5899
K-Means	1	20	1310	2	0,8044	1,0000	1,0000	1,0000	1,0000
Complete	1	20	1310	2	0,8044	1,0000	1,0000	1,0000	1,0000
Ward	1	20	1310	2	0,8044	1,0000	1,0000	1,0000	1,0000
K-Means	1	20	1320	2	0,7121	0,8126	0,5127	0,3531	0,4182
Complete	1	20	1320	2	0,7218	0,8086	0,0190	0,0407	0,0259
Ward	1	20	1320	2	0,7121	0,8126	0,5127	0,3531	0,4182
K-Means	1	20	1370	2	0,7741	0,8326	0,3297	0,0586	0,0995
Complete	1	20	1370	2	0,7249	0,9648	0,0035	0,0035	0,0035
Ward	1	20	1370	2	0,7741	0,8326	0,3297	0,0586	0,0995
K-Means	1	20	1380	2	0,9559	0,7516	0,1398	0,3449	0,1989

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Complete	1	20	1380	2	0,9559	0,7516	0,1398	0,3449	0,1989
Ward	1	20	1380	2	0,9559	0,7516	0,1398	0,3449	0,1989
K-Means	1	20	1390	2	0,9870	1,0000	1,0000	1,0000	1,0000
Complete	1	20	1390	2	0,9870	1,0000	1,0000	1,0000	1,0000
Ward	1	20	1390	2	0,9870	1,0000	1,0000	1,0000	1,0000
K-Means	1	20	1420	6	0,6738	0,6041	0,5570	0,5538	0,5554
Complete	1	20	1420	6	0,8904	0,7767	0,5194	0,8122	0,6336
Ward	1	20	1420	6	0,6864	0,6100	0,5590	0,5700	0,5645
K-Means	1	20	1430	2	0,7618	0,7709	0,4834	0,5610	0,5193
Complete	1	20	1430	2	0,7092	1,0000	1,0000	1,0000	1,0000
Ward	1	20	1430	2	0,7618	0,7709	0,4834	0,5610	0,5193
K-Means	1	20	1440	30	0,6136	0,5975	0,7960	0,8319	0,8135
Complete	1	20	1440	30	0,5406	0,5456	0,7645	0,8342	0,7978
Ward	1	20	1440	30	0,6262	0,6058	0,8184	0,8332	0,8257
K-Means	1	20	1480	4	0,6894	0,8721	0,7901	0,7643	0,7770
Complete	1	20	1480	4	0,7540	0,6906	0,3924	0,8113	0,5289
Ward	1	20	1480	4	0,6672	0,7331	0,6446	0,6371	0,6409
K-Means	1	20	1500	2	0,9750	1,0000	1,0000	1,0000	1,0000
Complete	1	20	1500	2	0,9750	1,0000	1,0000	1,0000	1,0000
Ward	1	20	1500	2	0,9750	1,0000	1,0000	1,0000	1,0000
K-Means	1	20	1530	2	0,7486	0,7236	0,4321	0,5024	0,4646
Complete	1	20	1530	2	0,7486	0,7236	0,4321	0,5024	0,4646
Ward	1	20	1530	2	0,7486	0,7236	0,4321	0,5024	0,4646
K-Means	1	20	1690	3	0,7341	0,7006	0,7337	0,6293	0,6775
Complete	1	20	1690	3	0,7341	0,7006	0,7337	0,6293	0,6775
Ward	1	20	1690	3	0,7341	0,7006	0,7337	0,6293	0,6775
K-Means	1	25	10	9	0,6807	0,3666	0,4113	0,1382	0,2068
Complete	1	25	10	9	0,4430	0,5028	0,4034	0,1767	0,2458
Ward	1	25	10	9	0,6532	0,3765	0,4285	0,1446	0,2162
K-Means	1	25	40	67	0,4498	0,1915	0,5547	0,3751	0,4476
Complete	1	25	40	67	0,4531	0,2075	0,4746	0,3735	0,4180
Ward	1	25	40	67	0,4276	0,2015	0,5554	0,3790	0,4506
K-Means	1	50	10	5	0,6028	0,4533	0,2249	0,1883	0,2050
Complete	1	50	10	5	0,5614	0,4233	0,1924	0,1556	0,1721
Ward	1	50	10	5	0,5898	0,4307	0,1255	0,1168	0,1210
K-Means	2	10	10	6	0,8371	0,5203	0,4390	0,3709	0,4021
Complete	2	10	10	6	0,8407	0,6248	0,3832	0,4430	0,4109
Ward	2	10	10	6	0,8366	0,5188	0,4333	0,3651	0,3963
K-Means	2	10	25	8	0,7130	0,7093	0,2639	0,0807	0,1236
Complete	2	10	25	8	0,8508	0,8967	0,1537	0,1202	0,1349
Ward	2	10	25	8	0,6856	0,6895	0,2795	0,0831	0,1281

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
K-Means	2	10	50	3	0,9116	0,9139	0,9115	0,5058	0,6506
Complete	2	10	50	3	0,9166	0,9304	0,9170	0,5463	0,6847
Ward	2	10	50	3	0,9116	0,9139	0,9115	0,5058	0,6506
K-Means	2	10	60	2	0,9464	0,9815	0,0023	0,0006	0,0010
Complete	2	10	60	2	0,9464	0,9815	0,0023	0,0006	0,0010
Ward	2	10	60	2	0,9464	0,9815	0,0023	0,0006	0,0010
K-Means	2	10	70	8	0,7193	0,6027	0,7152	0,1229	0,2097
Complete	2	10	70	8	0,7409	0,9263	0,7241	0,3356	0,4587
Ward	2	10	70	8	0,7194	0,6017	0,7214	0,1238	0,2113
K-Means	2	10	80	2	0,5981	0,7172	0,0817	0,0054	0,0101
Complete	2	10	80	2	0,5981	0,7172	0,0817	0,0054	0,0101
Ward	2	10	80	2	0,5981	0,7172	0,0817	0,0054	0,0101
K-Means	2	10	110	2	0,7669	0,7895	0,0509	0,0097	0,0163
Complete	2	10	110	2	0,7606	0,8212	0,0397	0,0086	0,0141
Ward	2	10	110	2	0,7026	0,7577	0,2473	0,0401	0,0691
K-Means	2	10	230	2	0,8803	1,0000	1,0000	1,0000	1,0000
Complete	2	10	230	2	0,9290	0,8026	0,0154	0,2385	0,0289
Ward	2	10	230	2	0,8803	1,0000	1,0000	1,0000	1,0000
K-Means	2	10	260	3	0,9399	0,9193	0,7831	0,3744	0,5066
Complete	2	10	260	3	0,9399	0,9193	0,7831	0,3744	0,5066
Ward	2	10	260	3	0,9399	0,9193	0,7831	0,3744	0,5066
K-Means	2	10	300	2	0,9974	1,0000	1,0000	1,0000	1,0000
Complete	2	10	300	2	0,9974	1,0000	1,0000	1,0000	1,0000
Ward	2	10	300	2	0,9974	1,0000	1,0000	1,0000	1,0000
K-Means	2	20	20	6	0,7094	0,6524	0,3920	0,3092	0,3457
Complete	2	20	20	6	0,7715	0,8020	0,4063	0,4982	0,4476
Ward	2	20	20	6	0,6612	0,6208	0,4115	0,2984	0,3459
K-Means	2	20	25	46	0,6115	0,2529	0,6702	0,5624	0,6116
Complete	2	20	25	46	0,5903	0,2619	0,6141	0,5672	0,5897
Ward	2	20	25	46	0,5979	0,3047	0,7033	0,5849	0,6387
K-Means	2	20	28	22	0,5729	0,4059	0,5314	0,5060	0,5184
Complete	2	20	28	22	0,5717	0,4798	0,4384	0,5917	0,5037
Ward	2	20	28	22	0,5670	0,4339	0,5334	0,5310	0,5322
K-Means	2	20	70	10	0,6212	0,5995	0,7923	0,2054	0,3263
Complete	2	20	70	10	0,8170	0,8828	0,8039	0,4115	0,5443
Ward	2	20	70	10	0,6207	0,5960	0,7936	0,2053	0,3262
K-Means	2	20	130	4	0,9739	1,0000	1,0000	1,0000	1,0000
Complete	2	20	130	4	0,9739	1,0000	1,0000	1,0000	1,0000
Ward	2	20	130	4	0,9739	1,0000	1,0000	1,0000	1,0000
K-Means	2	20	140	2	0,8806	0,9713	0,8560	0,7988	0,8264
Complete	2	20	140	2	0,8288	0,8261	0,0083	0,0369	0,0135

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Ward	2	20	140	2	0,8806	0,9713	0,8560	0,7988	0,8264
K-Means	2	20	150	4	0,5892	0,4435	0,5790	0,4649	0,5157
Complete	2	20	150	4	0,5482	0,5280	0,5501	0,4417	0,4900
Ward	2	20	150	4	0,6099	0,4566	0,5535	0,4638	0,5047
K-Means	2	20	210	2	0,7418	0,5988	0,0881	0,1285	0,1045
Complete	2	20	210	2	0,7418	0,5988	0,0881	0,1285	0,1045
Ward	2	20	210	2	0,7418	0,5988	0,0881	0,1285	0,1045
K-Means	2	20	220	2	0,8624	1,0000	1,0000	1,0000	1,0000
Complete	2	20	220	2	0,8624	1,0000	1,0000	1,0000	1,0000
Ward	2	20	220	2	0,8624	1,0000	1,0000	1,0000	1,0000
K-Means	2	20	230	2	0,8224	0,8430	0,5165	0,3175	0,3932
Complete	2	20	230	2	0,8224	0,8430	0,5165	0,3175	0,3932
Ward	2	20	230	2	0,8224	0,8430	0,5165	0,3175	0,3932
K-Means	2	30	29	17	0,5301	0,3703	0,4002	0,1192	0,1837
Complete	2	30	29	17	0,5717	0,4916	0,3342	0,1305	0,1878
Ward	2	30	29	17	0,5051	0,3641	0,3950	0,1173	0,1809
K-Means	2	30	30	84	0,3806	0,1979	0,4261	0,2706	0,3310
Complete	2	30	30	84	0,4178	0,3105	0,3666	0,3030	0,3318
Ward	2	30	30	84	0,3414	0,1922	0,4335	0,2728	0,3349
K-Means	2	30	31	7	0,6673	0,4662	0,5138	0,3744	0,4332
Complete	2	30	31	7	0,5998	0,5157	0,4546	0,4084	0,4302
Ward	2	30	31	7	0,6552	0,4662	0,5155	0,3745	0,4338
K-Means	2	30	36	3	0,9236	1,0000	1,0000	1,0000	1,0000
Complete	2	30	36	3	0,9236	1,0000	1,0000	1,0000	1,0000
Ward	2	30	36	3	0,9236	1,0000	1,0000	1,0000	1,0000
K-Means	2	30	40	2	0,5804	0,7135	0,0756	0,0016	0,0031
Complete	2	30	40	2	0,6962	0,9863	0,5834	0,1025	0,1743
Ward	2	30	40	2	0,5713	0,7059	0,0888	0,0018	0,0036
K-Means	2	30	42	6	0,5561	0,6923	0,3321	0,0303	0,0555
Complete	2	30	42	6	0,9117	0,9586	0,1528	0,0690	0,0951
Ward	2	30	42	6	0,5177	0,6589	0,3623	0,0314	0,0579
K-Means	2	30	110	6	0,5712	0,4344	0,3248	0,1637	0,2177
Complete	2	30	110	6	0,5566	0,5027	0,2447	0,1610	0,1942
Ward	2	30	110	6	0,5539	0,4441	0,3478	0,1740	0,2319
K-Means	2	30	130	7	0,7318	0,5957	0,6883	0,5189	0,5917
Complete	2	30	130	7	0,7294	0,5973	0,6816	0,5138	0,5859
Ward	2	30	130	7	0,7307	0,5937	0,6760	0,5089	0,5807
K-Means	2	30	140	5	0,5420	0,6498	0,8281	0,2742	0,4120
Complete	2	30	140	5	0,5288	0,7029	0,8775	0,3577	0,5082
Ward	2	30	140	5	0,5637	0,6462	0,8371	0,2785	0,4179
K-Means	2	30	170	4	0,7918	0,7159	0,5116	0,3995	0,4486

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Complete	2	30	170	4	0,7965	0,7310	0,5615	0,4495	0,4993
Ward	2	30	170	4	0,7965	0,7310	0,5615	0,4495	0,4993
K-Means	2	30	270	3	0,9117	0,7473	0,9240	0,5198	0,6653
Complete	2	30	270	3	0,9435	1,0000	1,0000	1,0000	1,0000
Ward	2	30	270	3	0,9117	0,7473	0,9240	0,5198	0,6653
K-Means	2	30	280	3	0,7881	0,8552	0,7153	0,7470	0,7308
Complete	2	30	280	3	0,8251	0,7020	0,3608	0,4702	0,4083
Ward	2	30	280	3	0,7881	0,8552	0,7153	0,7470	0,7308
K-Means	2	30	300	2	0,7407	1,0000	1,0000	1,0000	1,0000
Complete	2	30	300	2	0,7407	1,0000	1,0000	1,0000	1,0000
Ward	2	30	300	2	0,7407	1,0000	1,0000	1,0000	1,0000
K-Means	2	40	10	26	0,5445	0,3080	0,6029	0,1912	0,2903
Complete	2	40	10	26	0,5168	0,4548	0,5445	0,2254	0,3189
Ward	2	40	10	26	0,5262	0,3325	0,6128	0,1964	0,2974
K-Means	2	40	30	26	0,5734	0,2687	0,5556	0,3937	0,4609
Complete	2	40	30	26	0,5134	0,2866	0,5096	0,4068	0,4525
Ward	2	40	30	26	0,5610	0,2466	0,5543	0,3828	0,4529
K-Means	2	40	33	3	0,6142	0,7587	0,8968	0,2020	0,3297
Complete	2	40	33	3	0,3551	0,7624	0,8968	0,2036	0,3318
Ward	2	40	33	3	0,5777	0,7153	0,8968	0,1861	0,3082
K-Means	2	40	37	3	0,5360	0,5833	0,2606	0,0835	0,1264
Complete	2	40	37	3	0,5471	0,7429	0,2548	0,1154	0,1589
Ward	2	40	37	3	0,5361	0,5817	0,1708	0,0600	0,0888
K-Means	2	40	40	4	0,6700	0,4272	0,2027	0,1830	0,1924
Complete	2	40	40	4	0,9221	0,5426	0,1108	0,2131	0,1458
Ward	2	40	40	4	0,6765	0,4149	0,1941	0,1749	0,1840
K-Means	2	40	50	64	0,4465	0,2812	0,5066	0,3388	0,4060
Complete	2	40	50	64	0,4524	0,3755	0,3991	0,3807	0,3897
Ward	2	40	50	64	0,4192	0,2919	0,5168	0,3434	0,4126
K-Means	2	40	100	2	0,8684	0,8689	0,0003	0,0003	0,0003
Complete	2	40	100	2	0,9137	0,9175	0,0022	0,0138	0,0038
Ward	2	40	100	2	0,8684	0,8689	0,0003	0,0003	0,0003
K-Means	2	40	128	54	0,4375	0,2508	0,6775	0,3239	0,4383
Complete	2	40	128	54	0,4490	0,3482	0,5988	0,3562	0,4467
Ward	2	40	128	54	0,4234	0,2605	0,6755	0,3287	0,4423
K-Means	2	40	160	17	0,5343	0,5391	0,7335	0,4956	0,5915
Complete	2	40	160	17	0,5519	0,5614	0,6406	0,4818	0,5500
Ward	2	40	160	17	0,5413	0,5402	0,7371	0,5056	0,5998
K-Means	2	40	170	2	0,7792	1,0000	1,0000	1,0000	1,0000
Complete	2	40	170	2	0,7792	1,0000	1,0000	1,0000	1,0000
Ward	2	40	170	2	0,7792	1,0000	1,0000	1,0000	1,0000

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
K-Means	2	40	220	2	0,9344	0,9312	0,7944	0,7523	0,7728
Complete	2	40	220	2	0,9344	0,9312	0,7944	0,7523	0,7728
Ward	2	40	220	2	0,9344	0,9312	0,7944	0,7523	0,7728
K-Means	2	40	230	2	0,6883	0,7769	0,5924	0,5563	0,5738
Complete	2	40	230	2	0,6613	0,5484	0,2464	0,2641	0,2549
Ward	2	40	230	2	0,6883	0,7769	0,5924	0,5563	0,5738
K-Means	2	40	240	6	0,9101	0,7062	0,7029	0,3044	0,4248
Complete	2	40	240	6	0,9009	0,7052	0,6387	0,2825	0,3917
Ward	2	40	240	6	0,9090	0,7062	0,6951	0,3020	0,4211
K-Means	2	40	290	3	0,8897	1,0000	1,0000	1,0000	1,0000
Complete	2	40	290	3	0,8897	1,0000	1,0000	1,0000	1,0000
Ward	2	40	290	3	0,8897	1,0000	1,0000	1,0000	1,0000
K-Means	2	40	320	2	0,5846	0,6407	0,1806	0,0818	0,1126
Complete	2	40	320	2	0,7048	1,0000	1,0000	1,0000	1,0000
Ward	2	40	320	2	0,5846	0,6407	0,1806	0,0818	0,1126
K-Means	2	40	330	3	0,7256	0,9142	0,7484	0,7158	0,7317
Complete	2	40	330	3	0,7454	0,7113	0,2178	0,4229	0,2875
Ward	2	40	330	3	0,7256	0,9142	0,7484	0,7158	0,7317
K-Means	2	40	350	2	0,6349	0,3333	0,0205	0,0205	0,0205
Complete	2	40	350	2	0,6349	0,3333	0,0205	0,0205	0,0205
Ward	2	40	350	2	0,6349	0,3333	0,0205	0,0205	0,0205
K-Means	2	40	360	2	0,8927	1,0000	1,0000	1,0000	1,0000
Complete	2	40	360	2	0,8927	1,0000	1,0000	1,0000	1,0000
Ward	2	40	360	2	0,8927	1,0000	1,0000	1,0000	1,0000
K-Means	2	50	20	4	0,6205	0,3873	0,2892	0,2283	0,2552
Complete	2	50	20	4	0,6142	0,3837	0,2492	0,1999	0,2219
Ward	2	50	20	4	0,6046	0,3820	0,2886	0,2264	0,2537
K-Means	2	60	20	4	0,8606	0,5905	0,2525	0,1444	0,1837
Complete	2	60	20	4	0,8893	0,6339	0,2637	0,1794	0,2135
Ward	2	60	20	4	0,8618	0,5924	0,2525	0,1452	0,1843
K-Means	2	60	34	2	0,9267	0,9225	0,0094	0,0094	0,0094
Complete	2	60	34	2	0,9267	0,9225	0,0094	0,0094	0,0094
Ward	2	60	34	2	0,9267	0,9225	0,0094	0,0094	0,0094
K-Means	2	60	40	223	0,3547	0,1249	0,5324	0,2309	0,3221
Complete	2	60	40	223	0,3928	0,2475	0,4397	0,2589	0,3259
Ward	2	60	40	223	0,3227	0,1306	0,5408	0,2330	0,3257
K-Means	2	60	60	5	0,8197	0,7626	0,6780	0,8695	0,7619
Complete	2	60	60	5	0,9100	0,7592	0,6630	0,9060	0,7657
Ward	2	60	60	5	0,8197	0,7626	0,6780	0,8695	0,7619
K-Means	2	60	90	4	0,6316	0,7487	0,7906	0,4999	0,6125
Complete	2	60	90	4	0,7221	0,9807	0,9159	0,9057	0,9108

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Ward	2	60	90	4	0,6621	0,8208	0,8283	0,5786	0,6813
K-Means	2	60	98	3	0,6595	0,4000	0,5636	0,5636	0,5636
Complete	2	60	98	3	0,6595	0,4000	0,5636	0,5636	0,5636
Ward	2	60	98	3	0,6595	0,4000	0,5636	0,5636	0,5636
K-Means	2	60	120	99	0,4489	0,1597	0,4879	0,3188	0,3856
Complete	2	60	120	99	0,4273	0,1872	0,4342	0,3283	0,3739
Ward	2	60	120	99	0,4208	0,1640	0,4939	0,3231	0,3907
K-Means	2	60	175	2	0,9119	1,0000	1,0000	1,0000	1,0000
Complete	2	60	175	2	0,9119	1,0000	1,0000	1,0000	1,0000
Ward	2	60	175	2	0,9119	1,0000	1,0000	1,0000	1,0000
K-Means	2	60	200	6	0,6288	0,5833	0,6254	0,3602	0,4571
Complete	2	60	200	6	0,6268	0,5867	0,6302	0,3629	0,4606
Ward	2	60	200	6	0,6237	0,5908	0,6179	0,3618	0,4564
K-Means	2	60	220	3	0,8667	0,8260	0,7087	0,7628	0,7347
Complete	2	60	220	3	0,8667	0,8260	0,7087	0,7628	0,7347
Ward	2	60	220	3	0,8667	0,8260	0,7087	0,7628	0,7347
K-Means	2	60	240	3	0,8943	0,9438	0,8781	0,8176	0,8468
Complete	2	60	240	3	0,8943	0,9438	0,8781	0,8176	0,8468
Ward	2	60	240	3	0,8943	0,9438	0,8781	0,8176	0,8468
K-Means	2	60	260	4	0,7485	0,6585	0,6476	0,4283	0,5156
Complete	2	60	260	4	0,7500	0,6651	0,6514	0,4321	0,5196
Ward	2	60	260	4	0,7500	0,6651	0,6514	0,4321	0,5196
K-Means	2	60	270	4	0,7322	0,6862	0,5116	0,2130	0,3008
Complete	2	60	270	4	0,7258	0,6992	0,5116	0,2196	0,3073
Ward	2	60	270	4	0,7260	0,6820	0,5116	0,2110	0,2988
K-Means	2	60	450	2	0,5392	0,5744	0,1365	0,1391	0,1378
Complete	2	60	450	2	0,5392	0,5744	0,1365	0,1391	0,1378
Ward	2	60	450	2	0,5392	0,5744	0,1365	0,1391	0,1378
K-Means	2	60	460	2	0,7349	0,5455	0,1215	0,1215	0,1215
Complete	2	60	460	2	0,7349	0,5455	0,1215	0,1215	0,1215
Ward	2	60	460	2	0,7349	0,5455	0,1215	0,1215	0,1215
K-Means	2	70	20	4	0,8680	0,6890	0,6671	0,7577	0,7095
Complete	2	70	20	4	0,8680	0,6890	0,6671	0,7577	0,7095
Ward	2	70	20	4	0,8680	0,6890	0,6671	0,7577	0,7095
K-Means	2	70	40	2	0,7620	0,8332	0,0291	0,0011	0,0020
Complete	2	70	40	2	0,8497	0,9728	0,0036	0,0005	0,0009
Ward	2	70	40	2	0,8497	0,9728	0,0036	0,0005	0,0009
K-Means	2	70	50	6	0,6684	0,6273	0,6176	0,6002	0,6088
Complete	2	70	50	6	0,5872	0,6623	0,6057	0,6458	0,6251
Ward	2	70	50	6	0,6669	0,6217	0,6116	0,5973	0,6044
K-Means	2	70	98	6	0,7078	0,5755	0,7554	0,3655	0,4927

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Complete	2	70	98	6	0,5027	0,8146	0,7559	0,5008	0,6025
Ward	2	70	98	6	0,7062	0,5771	0,7563	0,3669	0,4941
K-Means	2	70	180	2	0,6602	1,0000	1,0000	1,0000	1,0000
Complete	2	70	180	2	0,6602	1,0000	1,0000	1,0000	1,0000
Ward	2	70	180	2	0,6602	1,0000	1,0000	1,0000	1,0000
K-Means	2	70	240	2	0,9267	0,9258	0,4086	0,0488	0,0872
Complete	2	70	240	2	0,9267	0,9258	0,4086	0,0488	0,0872
Ward	2	70	240	2	0,9267	0,9258	0,4086	0,0488	0,0872
K-Means	2	80	10	2	0,8768	0,9137	0,0142	0,0017	0,0031
Complete	2	80	10	2	0,8786	0,9183	0,0134	0,0017	0,0030
Ward	2	80	10	2	0,8786	0,9183	0,0134	0,0017	0,0030
K-Means	2	90	10	2	0,9322	0,7707	0,0414	0,0375	0,0394
Complete	2	90	10	2	0,9322	0,7707	0,0414	0,0375	0,0394
Ward	2	90	10	2	0,9322	0,7707	0,0414	0,0375	0,0394
K-Means	2	100	10	3	0,9035	0,9053	0,1195	0,0281	0,0455
Complete	2	100	10	3	0,9035	0,9053	0,1195	0,0281	0,0455
Ward	2	100	10	3	0,9035	0,9053	0,1195	0,0281	0,0455
K-Means	2	115	10	2	0,7797	0,8175	0,0383	0,0041	0,0074
Complete	2	115	10	2	0,7797	0,8175	0,0383	0,0041	0,0074
Ward	2	115	10	2	0,7681	0,7852	0,0488	0,0046	0,0085
K-Means	2	120	10	6	0,7358	0,6616	0,4527	0,4141	0,4325
Complete	2	120	10	6	0,5918	0,6970	0,4384	0,4163	0,4271
Ward	2	120	10	6	0,6455	0,7141	0,5292	0,4301	0,4745
K-Means	2	130	10	6	0,6209	0,5686	0,4712	0,2762	0,3483
Complete	2	130	10	6	0,5786	0,5740	0,4751	0,2917	0,3615
Ward	2	130	10	6	0,6147	0,5796	0,4956	0,2974	0,3717
K-Means	2	140	10	3	0,8206	0,9247	0,4701	0,4788	0,4745
Complete	2	140	10	3	0,8703	0,8690	0,0883	0,2778	0,1340
Ward	2	140	10	3	0,8221	0,9252	0,4855	0,5131	0,4989
K-Means	2	150	10	2	0,8589	0,6937	0,1162	0,0159	0,0280
Complete	2	150	10	2	0,8589	0,6937	0,1162	0,0159	0,0280
Ward	2	150	10	2	0,8589	0,6937	0,1162	0,0159	0,0280
K-Means	2	160	10	3	0,6905	0,5402	0,1509	0,0752	0,1004
Complete	2	160	10	3	0,7402	0,6663	0,2195	0,1632	0,1872
Ward	2	160	10	3	0,6915	0,5439	0,1473	0,0741	0,0986
K-Means	2	160	20	10	0,6548	0,3985	0,6204	0,5027	0,5554
Complete	2	160	20	10	0,6531	0,5102	0,6349	0,5424	0,5850
Ward	2	160	20	10	0,6545	0,3961	0,6097	0,4944	0,5460
K-Means	2	170	16	2	0,8996	0,9178	0,0138	0,0035	0,0056
Complete	2	170	16	2	0,8996	0,9178	0,0138	0,0035	0,0056
Ward	2	170	16	2	0,8996	0,9178	0,0138	0,0035	0,0056

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
K-Means	2	170	120	3	0,8715	0,7272	0,8681	0,4925	0,6285
Complete	2	170	120	3	0,8715	0,7272	0,8681	0,4925	0,6285
Ward	2	170	120	3	0,8715	0,7272	0,8681	0,4925	0,6285
K-Means	2	170	130	2	0,9911	1,0000	1,0000	1,0000	1,0000
Complete	2	170	130	2	0,9911	1,0000	1,0000	1,0000	1,0000
Ward	2	170	130	2	0,9911	1,0000	1,0000	1,0000	1,0000
K-Means	2	170	150	6	0,6589	0,6271	0,6439	0,5493	0,5929
Complete	2	170	150	6	0,5735	0,5969	0,6145	0,5130	0,5592
Ward	2	170	150	6	0,6560	0,6284	0,6498	0,5546	0,5984
K-Means	2	170	210	2	0,6890	0,6454	0,1308	0,0682	0,0897
Complete	2	170	210	2	0,6890	0,6454	0,1308	0,0682	0,0897
Ward	2	170	210	2	0,6890	0,6454	0,1308	0,0682	0,0897
K-Means	2	170	260	4	0,8608	0,5808	0,3036	0,3195	0,3113
Complete	2	170	260	4	0,8943	0,6374	0,3345	0,4358	0,3785
Ward	2	170	260	4	0,8582	0,5812	0,3171	0,3325	0,3246
K-Means	3	10	20	49	0,5053	0,4459	0,6203	0,2878	0,3932
Complete	3	10	20	49	0,4892	0,4658	0,5165	0,3162	0,3923
Ward	3	10	20	49	0,5013	0,4519	0,6411	0,3020	0,4105
K-Means	3	10	25	2	0,9147	1,0000	1,0000	1,0000	1,0000
Complete	3	10	25	2	0,9147	1,0000	1,0000	1,0000	1,0000
Ward	3	10	25	2	0,9147	1,0000	1,0000	1,0000	1,0000
K-Means	3	10	28	2	0,7032	0,7805	0,0475	0,0029	0,0054
Complete	3	10	28	2	0,7147	0,8222	0,0348	0,0024	0,0045
Ward	3	10	28	2	0,7147	0,8222	0,0348	0,0024	0,0045
K-Means	3	10	50	7	0,5666	0,6057	0,7958	0,3735	0,5084
Complete	3	10	50	7	0,7915	0,9359	0,7536	0,7017	0,7267
Ward	3	10	50	7	0,5778	0,7154	0,7801	0,4455	0,5671
K-Means	3	10	129	5	0,5657	0,6651	0,5338	0,1682	0,2558
Complete	3	10	129	5	0,8717	0,9393	0,5194	0,4301	0,4705
Ward	3	10	129	5	0,5536	0,7299	0,5408	0,1873	0,2783
K-Means	3	10	150	2	0,9291	0,9735	0,5844	0,1600	0,2512
Complete	3	10	150	2	0,9223	0,9670	0,5510	0,1313	0,2120
Ward	3	10	150	2	0,9291	0,9735	0,5844	0,1600	0,2512
K-Means	3	10	180	6	0,8281	0,6009	0,4122	0,0659	0,1136
Complete	3	10	180	6	0,8105	0,5989	0,3854	0,0614	0,1060
Ward	3	10	180	6	0,8240	0,5998	0,3996	0,0637	0,1099
K-Means	3	10	260	4	0,6226	0,7381	0,6323	0,5059	0,5621
Complete	3	10	260	4	0,5216	0,6365	0,6339	0,4959	0,5565
Ward	3	10	260	4	0,6040	0,7468	0,6789	0,5476	0,6062
K-Means	3	10	290	3	0,7316	0,6087	0,1528	0,1366	0,1443
Complete	3	10	290	3	0,8202	0,6997	0,1075	0,1559	0,1272

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Ward	3	10	290	3	0,7040	0,6086	0,1508	0,1351	0,1425
K-Means	3	10	310	7	0,6881	0,5459	0,5649	0,4637	0,5093
Complete	3	10	310	7	0,6338	0,6650	0,5600	0,4910	0,5233
Ward	3	10	310	7	0,6869	0,5497	0,5678	0,4682	0,5132
K-Means	3	10	330	6	0,7929	0,7710	0,7403	0,9057	0,8147
Complete	3	10	330	6	0,9030	0,5619	0,4285	0,8030	0,5588
Ward	3	10	330	6	0,7929	0,7710	0,7403	0,9057	0,8147
K-Means	3	10	360	2	0,6563	0,7182	0,3203	0,1477	0,2022
Complete	3	10	360	2	0,3131	0,8285	0,0268	0,0288	0,0277
Ward	3	10	360	2	0,6338	0,7455	0,3557	0,1697	0,2298
K-Means	3	10	380	2	0,6630	1,0000	1,0000	1,0000	1,0000
Complete	3	10	380	2	0,6630	1,0000	1,0000	1,0000	1,0000
Ward	3	10	380	2	0,6630	1,0000	1,0000	1,0000	1,0000
K-Means	3	10	430	2	0,9001	1,0000	1,0000	1,0000	1,0000
Complete	3	10	430	2	0,9001	1,0000	1,0000	1,0000	1,0000
Ward	3	10	430	2	0,9001	1,0000	1,0000	1,0000	1,0000
K-Means	3	10	450	47	0,5176	0,3947	0,6737	0,4796	0,5603
Complete	3	10	450	47	0,5041	0,4200	0,6257	0,5036	0,5580
Ward	3	10	450	47	0,4427	0,3116	0,6813	0,4635	0,5517
K-Means	3	10	460	2	0,5013	0,4082	0,3837	0,3113	0,3437
Complete	3	10	460	2	0,5013	0,4082	0,3837	0,3113	0,3437
Ward	3	10	460	2	0,5013	0,4082	0,3837	0,3113	0,3437
K-Means	3	10	490	2	0,6373	0,8194	0,3758	0,1204	0,1823
Complete	3	10	490	2	0,8276	0,9490	0,0025	0,0075	0,0038
Ward	3	10	490	2	0,6373	0,8194	0,3758	0,1204	0,1823
K-Means	3	10	510	2	0,7671	0,9488	0,8550	0,8565	0,8558
Complete	3	10	510	2	0,4248	0,6651	0,0519	0,1812	0,0807
Ward	3	10	510	2	0,7700	1,0000	1,0000	1,0000	1,0000
K-Means	3	10	520	2	0,7290	1,0000	1,0000	1,0000	1,0000
Complete	3	10	520	2	0,7290	1,0000	1,0000	1,0000	1,0000
Ward	3	10	520	2	0,7290	1,0000	1,0000	1,0000	1,0000
K-Means	3	10	580	2	0,8531	0,7461	0,0674	0,0299	0,0415
Complete	3	10	580	2	0,8130	0,8236	0,0348	0,0222	0,0271
Ward	3	10	580	2	0,8531	0,7461	0,0674	0,0299	0,0415
K-Means	3	10	620	2	0,8652	0,9007	0,0091	0,0154	0,0115
Complete	3	10	620	2	0,8652	0,9007	0,0091	0,0154	0,0115
Ward	3	10	620	2	0,8652	0,9007	0,0091	0,0154	0,0115
K-Means	3	15	10	5	0,7221	0,7005	0,6807	0,6299	0,6543
Complete	3	15	10	5	0,7221	0,7005	0,6807	0,6299	0,6543
Ward	3	15	10	5	0,6775	0,5798	0,5627	0,4992	0,5290

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
K-Means	3	20	20	34	0,4798	0,1684	0,3174	0,2043	0,2486
Complete	3	20	20	34	0,4487	0,1989	0,2912	0,2243	0,2534
Ward	3	20	20	34	0,4451	0,1647	0,3208	0,2103	0,2540
K-Means	3	20	70	2	0,6722	0,8331	0,3219	0,2146	0,2575
Complete	3	20	70	2	0,6722	0,8331	0,3219	0,2146	0,2575
Ward	3	20	70	2	0,6722	0,8331	0,3219	0,2146	0,2575
K-Means	3	20	90	2	0,8932	1,0000	1,0000	1,0000	1,0000
Complete	3	20	90	2	0,6704	0,7702	0,1861	0,3917	0,2523
Ward	3	20	90	2	0,8932	1,0000	1,0000	1,0000	1,0000
K-Means	3	20	170	2	0,9240	1,0000	1,0000	1,0000	1,0000
Complete	3	20	170	2	0,9240	1,0000	1,0000	1,0000	1,0000
Ward	3	20	170	2	0,9240	1,0000	1,0000	1,0000	1,0000
K-Means	3	30	9	9	0,6782	0,4524	0,3376	0,1583	0,2155
Complete	3	30	9	9	0,4667	0,4807	0,2695	0,1472	0,1904
Ward	3	30	9	9	0,6619	0,4446	0,3777	0,1756	0,2397
K-Means	3	30	10	2	0,7907	0,6561	0,1197	0,2444	0,1607
Complete	3	30	10	2	0,7584	0,9156	0,7777	0,7843	0,7810
Ward	3	30	10	2	0,7908	0,6570	0,1152	0,2409	0,1559
K-Means	3	30	30	16	0,5585	0,3903	0,6063	0,2796	0,3827
Complete	3	30	30	16	0,3769	0,5551	0,5150	0,3346	0,4057
Ward	3	30	30	16	0,5736	0,4041	0,5833	0,2780	0,3766
K-Means	3	30	40	17	0,4888	0,3623	0,3193	0,1083	0,1618
Complete	3	30	40	17	0,5703	0,5617	0,2511	0,1393	0,1792
Ward	3	30	40	17	0,4791	0,3952	0,3245	0,1159	0,1708
K-Means	3	30	43	2	0,8744	0,7149	0,1918	0,0269	0,0473
Complete	3	30	43	2	0,8744	0,7149	0,1918	0,0269	0,0473
Ward	3	30	43	2	0,8744	0,7149	0,1918	0,0269	0,0473
K-Means	3	30	50	3	0,6525	0,5809	0,1589	0,0104	0,0194
Complete	3	30	50	3	0,6622	0,6532	0,1422	0,0111	0,0205
Ward	3	30	50	3	0,6518	0,5794	0,1573	0,0102	0,0192
K-Means	3	30	54	2	0,9631	1,0000	1,0000	1,0000	1,0000
Complete	3	30	54	2	0,9631	1,0000	1,0000	1,0000	1,0000
Ward	3	30	54	2	0,9631	1,0000	1,0000	1,0000	1,0000
K-Means	3	30	56	7	0,7694	0,7222	0,7559	0,8321	0,7922
Complete	3	30	56	7	0,7694	0,7222	0,7559	0,8321	0,7922
Ward	3	30	56	7	0,7694	0,7222	0,7559	0,8321	0,7922
K-Means	3	30	60	16	0,5411	0,3121	0,3923	0,2600	0,3127
Complete	3	30	60	16	0,5108	0,4112	0,2248	0,2472	0,2354
Ward	3	30	60	16	0,5212	0,3094	0,3772	0,2503	0,3009
K-Means	3	30	70	177	0,3509	0,1322	0,5503	0,3554	0,4319

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Complete	3	30	70	177	0,3789	0,2258	0,4552	0,4003	0,4260
Ward	3	30	70	177	0,3166	0,1370	0,5557	0,3613	0,4379
K-Means	3	30	110	13	0,5688	0,4603	0,4826	0,6478	0,5531
Complete	3	30	110	13	0,8296	0,4512	0,3038	0,7980	0,4401
Ward	3	30	110	13	0,5644	0,4450	0,4779	0,6404	0,5474
K-Means	3	30	160	53	0,4532	0,3648	0,6571	0,3620	0,4668
Complete	3	30	160	53	0,4943	0,5469	0,5749	0,3896	0,4645
Ward	3	30	160	53	0,4126	0,3578	0,6621	0,3602	0,4666
K-Means	3	30	190	3	0,8197	0,9471	0,8699	0,7141	0,7843
Complete	3	30	190	3	0,7942	0,9727	0,8699	0,8616	0,8657
Ward	3	30	190	3	0,8162	0,9479	0,8699	0,7174	0,7863
K-Means	3	30	200	11	0,5405	0,3264	0,4846	0,2006	0,2837
Complete	3	30	200	11	0,5136	0,3957	0,5244	0,2304	0,3201
Ward	3	30	200	11	0,5341	0,3327	0,5026	0,2060	0,2922
K-Means	3	30	230	10	0,7661	0,6824	0,6268	0,4860	0,5475
Complete	3	30	230	10	0,7700	0,6775	0,6115	0,5070	0,5544
Ward	3	30	230	10	0,7571	0,6826	0,6295	0,4883	0,5500
K-Means	3	30	240	4	0,5957	0,6514	0,3397	0,4096	0,3714
Complete	3	30	240	4	0,6757	0,6484	0,2053	0,3601	0,2615
Ward	3	30	240	4	0,6757	0,6484	0,2053	0,3601	0,2615
K-Means	3	30	250	2	0,9342	0,6755	0,0396	0,1269	0,0604
Complete	3	30	250	2	0,9342	0,6755	0,0396	0,1269	0,0604
Ward	3	30	250	2	0,9342	0,6755	0,0396	0,1269	0,0604
K-Means	3	30	300	19	0,6800	0,5585	0,5947	0,4865	0,5352
Complete	3	30	300	19	0,7031	0,5482	0,5557	0,5135	0,5338
Ward	3	30	300	19	0,6942	0,5618	0,5990	0,4971	0,5433
K-Means	3	30	310	13	0,6975	0,4836	0,5767	0,5047	0,5383
Complete	3	30	310	13	0,5128	0,4280	0,5102	0,5166	0,5134
Ward	3	30	310	13	0,7227	0,4677	0,5291	0,5038	0,5161
K-Means	3	30	360	6	0,5884	0,5990	0,5269	0,1120	0,1847
Complete	3	30	360	6	0,5777	0,7786	0,5654	0,1753	0,2677
Ward	3	30	360	6	0,5801	0,6042	0,5342	0,1156	0,1900
K-Means	3	30	365	2	0,9130	0,9236	0,0113	0,0067	0,0084
Complete	3	30	365	2	0,8743	0,9616	0,0028	0,0048	0,0035
Ward	3	30	365	2	0,9130	0,9236	0,0113	0,0067	0,0084
K-Means	3	30	370	2	0,9862	1,0000	1,0000	1,0000	1,0000
Complete	3	30	370	2	0,9862	1,0000	1,0000	1,0000	1,0000
Ward	3	30	370	2	0,9862	1,0000	1,0000	1,0000	1,0000
K-Means	3	30	380	2	0,6997	0,8972	0,4728	0,1603	0,2394
Complete	3	30	380	2	0,7197	0,9662	0,0011	0,0047	0,0018
Ward	3	30	380	2	0,6997	0,8972	0,4728	0,1603	0,2394

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
K-Means	3	30	390	7	0,6331	0,5789	0,7985	0,3477	0,4844
Complete	3	30	390	7	0,6474	0,6467	0,8323	0,4101	0,5495
Ward	3	30	390	7	0,6345	0,5919	0,8004	0,3554	0,4923
K-Means	3	30	410	2	0,7420	0,6424	0,1113	0,2292	0,1498
Complete	3	30	410	2	0,7721	0,6566	0,0721	0,1967	0,1055
Ward	3	30	410	2	0,7721	0,6566	0,0721	0,1967	0,1055
K-Means	3	30	420	25	0,5981	0,3612	0,6537	0,4458	0,5301
Complete	3	30	420	25	0,5721	0,3668	0,5810	0,4414	0,5017
Ward	3	30	420	25	0,5572	0,3540	0,6545	0,4410	0,5270
K-Means	3	30	428	5	0,6057	0,4968	0,5570	0,2604	0,3549
Complete	3	30	428	5	0,7503	0,6486	0,5216	0,3301	0,4044
Ward	3	30	428	5	0,6167	0,5330	0,5472	0,2713	0,3627
K-Means	3	30	430	2	0,6067	1,0000	1,0000	1,0000	1,0000
Complete	3	30	430	2	0,3677	0,6032	0,2706	0,1601	0,2012
Ward	3	30	430	2	0,6067	1,0000	1,0000	1,0000	1,0000
K-Means	3	30	450	26	0,5225	0,2891	0,4489	0,3009	0,3603
Complete	3	30	450	26	0,5056	0,3774	0,4361	0,3332	0,3778
Ward	3	30	450	26	0,5078	0,2535	0,4500	0,2918	0,3540
K-Means	3	30	457	2	0,9818	1,0000	1,0000	1,0000	1,0000
Complete	3	30	457	2	0,9818	1,0000	1,0000	1,0000	1,0000
Ward	3	30	457	2	0,9818	1,0000	1,0000	1,0000	1,0000
K-Means	3	30	460	5	0,7797	0,7340	0,7560	0,4964	0,5993
Complete	3	30	460	5	0,7859	0,7296	0,6871	0,4712	0,5590
Ward	3	30	460	5	0,7741	0,7257	0,7392	0,4907	0,5899
K-Means	3	30	465	2	0,9731	1,0000	1,0000	1,0000	1,0000
Complete	3	30	465	2	0,9731	1,0000	1,0000	1,0000	1,0000
Ward	3	30	465	2	0,9731	1,0000	1,0000	1,0000	1,0000
K-Means	3	30	470	9	0,5797	0,3692	0,3060	0,1037	0,1549
Complete	3	30	470	9	0,5266	0,4233	0,3110	0,1174	0,1704
Ward	3	30	470	9	0,5499	0,3615	0,2641	0,0875	0,1315
K-Means	3	30	479	2	0,7465	0,6362	0,0843	0,1650	0,1116
Complete	3	30	479	2	0,7465	0,6362	0,0843	0,1650	0,1116
Ward	3	30	479	2	0,7465	0,6362	0,0843	0,1650	0,1116
K-Means	3	30	500	2	0,9527	1,0000	1,0000	1,0000	1,0000
Complete	3	30	500	2	0,9527	1,0000	1,0000	1,0000	1,0000
Ward	3	30	500	2	0,9527	1,0000	1,0000	1,0000	1,0000
K-Means	3	30	505	3	0,8021	0,9031	0,0148	0,0060	0,0085
Complete	3	30	505	3	0,8432	0,9312	0,0089	0,0051	0,0065
Ward	3	30	505	3	0,8432	0,9312	0,0089	0,0051	0,0065
K-Means	3	30	530	4	0,5640	0,5960	0,1274	0,0118	0,0216
Complete	3	30	530	4	0,5281	0,7287	0,0823	0,0116	0,0203

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Ward	3	30	530	4	0,5488	0,6314	0,1490	0,0145	0,0264
K-Means	3	30	540	2	0,7413	0,7509	0,2221	0,0268	0,0479
Complete	3	30	540	2	0,7979	1,0000	1,0000	1,0000	1,0000
Ward	3	30	540	2	0,7242	0,7352	0,2037	0,0238	0,0426
K-Means	3	30	559	2	0,8232	0,6973	0,3334	0,1890	0,2413
Complete	3	30	559	2	0,8232	0,6973	0,3334	0,1890	0,2413
Ward	3	30	559	2	0,8232	0,6973	0,3334	0,1890	0,2413
K-Means	3	30	565	2	0,7317	0,7982	0,0125	0,0053	0,0075
Complete	3	30	565	2	0,8432	0,9485	0,0008	0,0075	0,0015
Ward	3	30	565	2	0,7383	0,8214	0,0169	0,0082	0,0110
K-Means	3	30	700	3	0,8151	0,8212	0,8282	0,3586	0,5005
Complete	3	30	700	3	0,8151	0,8212	0,8282	0,3586	0,5005
Ward	3	30	700	3	0,8151	0,8212	0,8282	0,3586	0,5005
K-Means	3	30	720	9	0,6278	0,6136	0,6892	0,6422	0,6649
Complete	3	30	720	9	0,6078	0,6714	0,6978	0,6945	0,6961
Ward	3	30	720	9	0,6178	0,6121	0,6892	0,6245	0,6553
K-Means	3	30	730	3	0,7208	0,9497	0,8371	0,8382	0,8377
Complete	3	30	730	3	0,6954	0,7622	0,3928	0,7086	0,5054
Ward	3	30	730	3	0,7208	0,9497	0,8371	0,8382	0,8377
K-Means	3	30	750	9	0,6122	0,5951	0,6159	0,3980	0,4835
Complete	3	30	750	9	0,7439	0,8306	0,6131	0,6117	0,6124
Ward	3	30	750	9	0,6377	0,6340	0,6292	0,4117	0,4977
K-Means	3	30	780	3	0,9563	0,8505	0,2064	0,8811	0,3344
Complete	3	30	780	3	0,9563	0,8505	0,2064	0,8811	0,3344
Ward	3	30	780	3	0,9563	0,8505	0,2064	0,8811	0,3344
K-Means	3	30	860	2	0,8933	1,0000	1,0000	1,0000	1,0000
Complete	3	30	860	2	0,9349	0,9859	0,1369	0,6064	0,2234
Ward	3	30	860	2	0,8933	1,0000	1,0000	1,0000	1,0000
K-Means	3	30	870	2	0,8613	0,7931	0,3973	0,1794	0,2472
Complete	3	30	870	2	0,8613	0,7931	0,3973	0,1794	0,2472
Ward	3	30	870	2	0,8613	0,7931	0,3973	0,1794	0,2472
K-Means	3	30	900	2	0,8085	0,7826	0,0528	0,0125	0,0203
Complete	3	30	900	2	0,8085	0,7826	0,0528	0,0125	0,0203
Ward	3	30	900	2	0,8085	0,7826	0,0528	0,0125	0,0203
K-Means	3	30	910	4	0,8733	0,8794	0,6418	0,7658	0,6983
Complete	3	30	910	4	0,8733	0,8794	0,6418	0,7658	0,6983
Ward	3	30	910	4	0,8733	0,8794	0,6418	0,7658	0,6983
K-Means	3	30	920	8	0,7381	0,4758	0,7321	0,4985	0,5932
Complete	3	30	920	8	0,7723	0,5412	0,7169	0,5484	0,6215
Ward	3	30	920	8	0,7381	0,4758	0,7321	0,4985	0,5932
K-Means	3	30	930	3	0,5808	0,6059	0,2075	0,0288	0,0506

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Complete	3	30	930	3	0,5519	0,6493	0,2075	0,0317	0,0551
Ward	3	30	930	3	0,5436	0,6049	0,3261	0,0439	0,0774
K-Means	3	30	950	2	0,8406	0,9567	0,7397	0,5755	0,6473
Complete	3	30	950	2	0,8492	0,9854	0,8710	0,7909	0,8290
Ward	3	30	950	2	0,8492	0,9854	0,8710	0,7909	0,8290
K-Means	3	30	980	3	0,9277	1,0000	1,0000	1,0000	1,0000
Complete	3	30	980	3	0,9277	1,0000	1,0000	1,0000	1,0000
Ward	3	30	980	3	0,9277	1,0000	1,0000	1,0000	1,0000
K-Means	3	30	1020	2	0,7763	1,0000	1,0000	1,0000	1,0000
Complete	3	30	1020	2	0,7763	1,0000	1,0000	1,0000	1,0000
Ward	3	30	1020	2	0,7763	1,0000	1,0000	1,0000	1,0000
K-Means	3	30	1040	4	0,8904	1,0000	1,0000	1,0000	1,0000
Complete	3	30	1040	4	0,8781	0,9924	0,8693	0,9270	0,8972
Ward	3	30	1040	4	0,8904	1,0000	1,0000	1,0000	1,0000
K-Means	3	30	1050	3	0,8306	0,7046	0,2575	0,0907	0,1341
Complete	3	30	1050	3	0,8306	0,7046	0,2575	0,0907	0,1341
Ward	3	30	1050	3	0,8306	0,7046	0,2575	0,0907	0,1341
K-Means	3	30	1070	2	0,9433	0,5976	0,1611	0,2193	0,1857
Complete	3	30	1070	2	0,9433	0,5976	0,1611	0,2193	0,1857
Ward	3	30	1070	2	0,9433	0,5976	0,1611	0,2193	0,1857
K-Means	3	30	1110	2	0,8163	1,0000	1,0000	1,0000	1,0000
Complete	3	30	1110	2	0,8100	0,9732	0,8654	0,8141	0,8389
Ward	3	30	1110	2	0,8163	1,0000	1,0000	1,0000	1,0000
K-Means	3	30	1120	7	0,7533	0,7339	0,8079	0,6465	0,7182
Complete	3	30	1120	7	0,7571	0,7417	0,8151	0,6540	0,7258
Ward	3	30	1120	7	0,7434	0,7486	0,8373	0,6729	0,7462
K-Means	3	30	1140	2	0,8756	0,8015	0,0447	0,0114	0,0181
Complete	3	30	1140	2	0,7996	1,0000	1,0000	1,0000	1,0000
Ward	3	30	1140	2	0,8756	0,8015	0,0447	0,0114	0,0181
K-Means	3	30	1180	2	0,7943	1,0000	1,0000	1,0000	1,0000
Complete	3	30	1180	2	0,7943	1,0000	1,0000	1,0000	1,0000
Ward	3	30	1180	2	0,7943	1,0000	1,0000	1,0000	1,0000
K-Means	3	30	1230	2	0,8747	0,8933	0,4277	0,1017	0,1643
Complete	3	30	1230	2	0,8428	1,0000	1,0000	1,0000	1,0000
Ward	3	30	1230	2	0,8747	0,8933	0,4277	0,1017	0,1643
K-Means	3	30	1240	2	0,6382	0,7508	0,4253	0,3907	0,4072
Complete	3	30	1240	2	0,5890	0,6593	0,3987	0,3648	0,3810
Ward	3	30	1240	2	0,5890	0,6593	0,3987	0,3648	0,3810
K-Means	3	30	1300	7	0,7030	0,7686	0,9240	0,7344	0,8183
Complete	3	30	1300	7	0,5849	0,6113	0,7015	0,6161	0,6560
Ward	3	30	1300	7	0,6450	0,7329	0,9240	0,6991	0,7960

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
K-Means	3	30	1310	3	0,4675	0,7485	0,7108	0,4960	0,5843
Complete	3	30	1310	3	0,4675	0,7485	0,7108	0,4960	0,5843
Ward	3	30	1310	3	0,4675	0,7485	0,7108	0,4960	0,5843
K-Means	3	30	1330	17	0,5284	0,3495	0,4706	0,3907	0,4269
Complete	3	30	1330	17	0,5017	0,4138	0,4361	0,4538	0,4448
Ward	3	30	1330	17	0,5231	0,3319	0,4576	0,3875	0,4197
K-Means	3	30	1360	16	0,7361	0,5014	0,5802	0,6923	0,6313
Complete	3	30	1360	16	0,6580	0,4767	0,5036	0,7765	0,6110
Ward	3	30	1360	16	0,7311	0,5095	0,5824	0,7029	0,6370
K-Means	3	30	1370	9	0,6570	0,6150	0,5767	0,4599	0,5117
Complete	3	30	1370	9	0,6665	0,6098	0,4816	0,4755	0,4785
Ward	3	30	1370	9	0,6469	0,6246	0,5885	0,4764	0,5266
K-Means	3	30	1380	2	0,6413	0,5813	0,1968	0,2110	0,2037
Complete	3	30	1380	2	0,6345	0,5473	0,0570	0,0533	0,0551
Ward	3	30	1380	2	0,6345	0,5473	0,0570	0,0533	0,0551
K-Means	3	30	1390	2	0,8682	0,9908	0,9397	0,9397	0,9397
Complete	3	30	1390	2	0,8322	0,8818	0,5054	0,6354	0,5630
Ward	3	30	1390	2	0,8679	0,9954	0,9724	0,9673	0,9699
K-Means	3	30	1460	7	0,6414	0,5000	0,5790	0,5685	0,5737
Complete	3	30	1460	7	0,6280	0,5378	0,5516	0,5888	0,5696
Ward	3	30	1460	7	0,6384	0,4896	0,6063	0,5818	0,5938
K-Means	3	30	1490	19	0,6529	0,5209	0,7618	0,7458	0,7537
Complete	3	30	1490	19	0,7048	0,5254	0,6709	0,7931	0,7269
Ward	3	30	1490	19	0,7121	0,4982	0,7290	0,7610	0,7446
K-Means	3	30	1610	2	0,8749	1,0000	1,0000	1,0000	1,0000
Complete	3	30	1610	2	0,8726	0,9466	0,5478	0,7171	0,6211
Ward	3	30	1610	2	0,8749	1,0000	1,0000	1,0000	1,0000
K-Means	3	30	1660	3	0,7264	0,5426	0,2909	0,1578	0,2046
Complete	3	30	1660	3	0,7111	0,5874	0,3869	0,2129	0,2746
Ward	3	30	1660	3	0,7204	0,5436	0,2988	0,1609	0,2092
K-Means	3	30	1680	3	0,6839	0,8078	0,7434	0,7559	0,7496
Complete	3	30	1680	3	0,6069	0,6756	0,5731	0,6699	0,6177
Ward	3	30	1680	3	0,6333	0,6702	0,5731	0,6642	0,6153
K-Means	3	30	1740	2	0,8863	0,6254	0,1323	0,1057	0,1175
Complete	3	30	1740	2	0,8863	0,6254	0,1323	0,1057	0,1175
Ward	3	30	1740	2	0,8863	0,6254	0,1323	0,1057	0,1175
K-Means	3	30	1990	2	0,6629	1,0000	1,0000	1,0000	1,0000
Complete	3	30	1990	2	0,6629	1,0000	1,0000	1,0000	1,0000
Ward	3	30	1990	2	0,6629	1,0000	1,0000	1,0000	1,0000
K-Means	3	30	2000	4	0,8151	0,5867	0,4099	0,6881	0,5138
Complete	3	30	2000	4	0,5091	0,5772	0,5301	0,5859	0,5566

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Ward	3	30	2000	4	0,8151	0,5867	0,4099	0,6881	0,5138
K-Means	3	30	2010	2	0,9772	0,5896	0,2449	0,1653	0,1974
Complete	3	30	2010	2	0,9772	0,5896	0,2449	0,1653	0,1974
Ward	3	30	2010	2	0,9772	0,5896	0,2449	0,1653	0,1974
K-Means	3	30	2020	2	0,9606	0,9496	0,8506	0,8584	0,8545
Complete	3	30	2020	2	0,9606	0,9496	0,8506	0,8584	0,8545
Ward	3	30	2020	2	0,9606	0,9496	0,8506	0,8584	0,8545
K-Means	3	30	2030	2	0,6536	1,0000	1,0000	1,0000	1,0000
Complete	3	30	2030	2	0,6536	1,0000	1,0000	1,0000	1,0000
Ward	3	30	2030	2	0,6536	1,0000	1,0000	1,0000	1,0000
K-Means	3	30	2120	2	0,9473	1,0000	1,0000	1,0000	1,0000
Complete	3	30	2120	2	0,9473	1,0000	1,0000	1,0000	1,0000
Ward	3	30	2120	2	0,9473	1,0000	1,0000	1,0000	1,0000
K-Means	3	30	2130	2	0,8210	0,6825	0,0561	0,0951	0,0706
Complete	3	30	2130	2	0,8210	0,6825	0,0561	0,0951	0,0706
Ward	3	30	2130	2	0,8210	0,6825	0,0561	0,0951	0,0706
K-Means	3	30	2220	3	0,9339	1,0000	1,0000	1,0000	1,0000
Complete	3	30	2220	3	0,9339	1,0000	1,0000	1,0000	1,0000
Ward	3	30	2220	3	0,9339	1,0000	1,0000	1,0000	1,0000
K-Means	3	30	2310	4	0,6300	0,7461	0,3962	0,1874	0,2545
Complete	3	30	2310	4	0,6314	0,9150	0,3726	0,3949	0,3834
Ward	3	30	2310	4	0,6250	0,7451	0,4450	0,2065	0,2821
K-Means	3	30	2320	5	0,7043	0,7113	0,7274	0,6864	0,7063
Complete	3	30	2320	5	0,7012	0,7075	0,7274	0,6844	0,7053
Ward	3	30	2320	5	0,7012	0,7075	0,7274	0,6844	0,7053
K-Means	3	30	2450	2	0,9222	1,0000	1,0000	1,0000	1,0000
Complete	3	30	2450	2	0,9222	1,0000	1,0000	1,0000	1,0000
Ward	3	30	2450	2	0,9222	1,0000	1,0000	1,0000	1,0000
K-Means	3	40	5	5	0,6367	0,7694	0,6361	0,8544	0,7292
Complete	3	40	5	5	0,5093	0,4961	0,2970	0,5326	0,3813
Ward	3	40	5	5	0,6367	0,7694	0,6361	0,8544	0,7292
K-Means	3	40	20	2	0,9412	0,9810	0,0020	0,0014	0,0016
Complete	3	40	20	2	0,9412	0,9810	0,0020	0,0014	0,0016
Ward	3	40	20	2	0,9412	0,9810	0,0020	0,0014	0,0016
K-Means	3	40	30	6	0,5045	0,5122	0,3684	0,0254	0,0475
Complete	3	40	30	6	0,4910	0,9525	0,3549	0,1494	0,2103
Ward	3	40	30	6	0,5812	0,6215	0,3858	0,0336	0,0619
K-Means	3	40	47	2	0,5940	0,7640	0,2024	0,0127	0,0239
Complete	3	40	47	2	0,8613	0,9927	0,0002	0,0008	0,0003
Ward	3	40	47	2	0,5831	0,8572	0,3036	0,0262	0,0482
K-Means	3	40	50	178	0,3479	0,0916	0,4247	0,2889	0,3439

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Complete	3	40	50	178	0,3097	0,1168	0,3848	0,3001	0,3372
Ward	3	40	50	178	0,2983	0,0888	0,4218	0,2879	0,3423
K-Means	3	40	91	8	0,7399	0,6703	0,7255	0,6640	0,6934
Complete	3	40	91	8	0,7254	0,6977	0,6933	0,6914	0,6923
Ward	3	40	91	8	0,7350	0,6715	0,7234	0,6765	0,6991
K-Means	3	40	109	2	0,9402	0,9857	0,0012	0,0012	0,0012
Complete	3	40	109	2	0,9538	0,9893	0,0006	0,0011	0,0008
Ward	3	40	109	2	0,9402	0,9857	0,0012	0,0012	0,0012
K-Means	3	40	140	3	0,6721	0,6380	0,2318	0,0125	0,0237
Complete	3	40	140	3	0,6716	0,6389	0,2315	0,0125	0,0237
Ward	3	40	140	3	0,6560	0,6215	0,2385	0,0125	0,0238
K-Means	3	40	190	7	0,6094	0,4780	0,5160	0,1460	0,2276
Complete	3	40	190	7	0,5373	0,5250	0,1978	0,0711	0,1046
Ward	3	40	190	7	0,5659	0,4596	0,3924	0,1067	0,1678
K-Means	3	40	220	2	0,9339	0,6234	0,2141	0,1285	0,1606
Complete	3	40	220	2	0,9339	0,6234	0,2141	0,1285	0,1606
Ward	3	40	220	2	0,9327	0,6229	0,2118	0,1273	0,1590
K-Means	3	40	250	2	0,8545	0,8833	0,0185	0,0008	0,0015
Complete	3	40	250	2	0,8502	0,9468	0,0076	0,0006	0,0011
Ward	3	40	250	2	0,8545	0,8833	0,0185	0,0008	0,0015
K-Means	3	40	366	2	0,9249	1,0000	1,0000	1,0000	1,0000
Complete	3	40	366	2	0,9249	1,0000	1,0000	1,0000	1,0000
Ward	3	40	366	2	0,9249	1,0000	1,0000	1,0000	1,0000
K-Means	3	40	390	4	0,6854	0,7929	0,5323	0,1788	0,2677
Complete	3	40	390	4	0,7428	0,7490	0,0623	0,0265	0,0372
Ward	3	40	390	4	0,6866	0,7675	0,5968	0,1868	0,2845
K-Means	3	40	600	3	0,5979	0,7907	0,0428	0,0096	0,0157
Complete	3	40	600	3	0,8372	0,9683	0,3644	0,3644	0,3644
Ward	3	40	600	3	0,5979	0,7907	0,0428	0,0096	0,0157
K-Means	3	40	630	8	0,6228	0,5399	0,6290	0,3660	0,4627
Complete	3	40	630	8	0,5529	0,5943	0,6148	0,3857	0,4741
Ward	3	40	630	8	0,6167	0,5361	0,6158	0,3611	0,4552
K-Means	3	40	710	2	0,7528	0,7190	0,1091	0,0009	0,0018
Complete	3	40	710	2	0,7483	0,7240	0,1144	0,0010	0,0020
Ward	3	40	710	2	0,7483	0,7240	0,1144	0,0010	0,0020
K-Means	3	40	800	2	0,7025	0,9887	0,9541	0,9555	0,9548
Complete	3	40	800	2	0,9133	0,7065	0,0048	0,1209	0,0093
Ward	3	40	800	2	0,7026	0,9943	0,9717	0,9717	0,9717
K-Means	3	40	850	2	0,9320	0,9682	0,0010	0,0044	0,0016
Complete	3	40	850	2	0,9320	0,9682	0,0010	0,0044	0,0016
Ward	3	40	850	2	0,9320	0,9682	0,0010	0,0044	0,0016

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
K-Means	3	40	970	3	0,9418	1,0000	1,0000	1,0000	1,0000
Complete	3	40	970	3	0,9418	1,0000	1,0000	1,0000	1,0000
Ward	3	40	970	3	0,9418	1,0000	1,0000	1,0000	1,0000
K-Means	3	40	1000	7	0,6186	0,5594	0,6618	0,4486	0,5348
Complete	3	40	1000	7	0,5388	0,5074	0,6868	0,4571	0,5489
Ward	3	40	1000	7	0,5962	0,5511	0,6848	0,4779	0,5629
K-Means	3	40	1190	3	0,5112	0,7134	0,8915	0,3016	0,4507
Complete	3	40	1190	3	0,5034	0,7937	0,8185	0,3230	0,4632
Ward	3	40	1190	3	0,5126	0,7183	0,8919	0,3039	0,4533
K-Means	3	40	1260	2	0,9191	1,0000	1,0000	1,0000	1,0000
Complete	3	40	1260	2	0,9191	1,0000	1,0000	1,0000	1,0000
Ward	3	40	1260	2	0,9191	1,0000	1,0000	1,0000	1,0000
K-Means	3	40	1280	2	0,5948	0,7253	0,1340	0,0589	0,0818
Complete	3	40	1280	2	0,5269	0,6520	0,1537	0,0635	0,0899
Ward	3	40	1280	2	0,6074	0,7450	0,0000	0,0000	0,0000
K-Means	3	40	1310	3	0,7604	0,5590	0,2538	0,0797	0,1213
Complete	3	40	1310	3	0,7685	0,7740	0,2246	0,1463	0,1772
Ward	3	40	1310	3	0,7568	0,5597	0,2569	0,0806	0,1227
K-Means	3	40	1350	11	0,6428	0,5468	0,7209	0,3927	0,5084
Complete	3	40	1350	11	0,6379	0,5949	0,7304	0,4154	0,5296
Ward	3	40	1350	11	0,6358	0,5017	0,7212	0,3818	0,4993
K-Means	3	40	1360	2	0,8328	0,6908	0,0301	0,0926	0,0454
Complete	3	40	1360	2	0,8328	0,6908	0,0301	0,0926	0,0454
Ward	3	40	1360	2	0,8328	0,6908	0,0301	0,0926	0,0454
K-Means	3	40	1380	2	0,5943	0,7171	0,3449	0,4872	0,4039
Complete	3	40	1380	2	0,5943	0,7171	0,3449	0,4872	0,4039
Ward	3	40	1380	2	0,5943	0,7171	0,3449	0,4872	0,4039
K-Means	3	40	1390	2	0,9678	1,0000	1,0000	1,0000	1,0000
Complete	3	40	1390	2	0,9678	1,0000	1,0000	1,0000	1,0000
Ward	3	40	1390	2	0,9678	1,0000	1,0000	1,0000	1,0000
K-Means	3	40	1420	3	0,6703	0,6113	0,3972	0,5283	0,4535
Complete	3	40	1420	3	0,6703	0,6113	0,3972	0,5283	0,4535
Ward	3	40	1420	3	0,6703	0,6113	0,3972	0,5283	0,4535
K-Means	3	40	1440	3	0,8780	0,6918	0,6320	0,3429	0,4446
Complete	3	40	1440	3	0,8780	0,6918	0,6320	0,3429	0,4446
Ward	3	40	1440	3	0,8780	0,6918	0,6320	0,3429	0,4446
K-Means	3	40	1500	2	0,5821	0,6522	0,2574	0,2280	0,2418
Complete	3	40	1500	2	0,5741	0,8489	0,5267	0,5115	0,5190
Ward	3	40	1500	2	0,5636	0,6026	0,1928	0,1711	0,1813
K-Means	3	40	1520	2	0,8954	1,0000	1,0000	1,0000	1,0000
Complete	3	40	1520	2	0,8954	1,0000	1,0000	1,0000	1,0000

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Ward	3	40	1520	2	0,8954	1,0000	1,0000	1,0000	1,0000
K-Means	3	40	1530	2	0,7800	0,6548	0,2195	0,0896	0,1272
Complete	3	40	1530	2	0,6928	1,0000	1,0000	1,0000	1,0000
Ward	3	40	1530	2	0,7800	0,6548	0,2195	0,0896	0,1272
K-Means	3	40	1570	2	0,9308	0,6069	0,1699	0,1015	0,1271
Complete	3	40	1570	2	0,9308	0,6069	0,1699	0,1015	0,1271
Ward	3	40	1570	2	0,9308	0,6069	0,1699	0,1015	0,1271
K-Means	3	40	1580	3	0,7177	0,7950	0,8028	0,7472	0,7740
Complete	3	40	1580	3	0,7177	0,7950	0,8028	0,7472	0,7740
Ward	3	40	1580	3	0,7177	0,7950	0,8028	0,7472	0,7740
K-Means	3	40	1620	3	0,8270	0,9080	0,8424	0,6952	0,7618
Complete	3	40	1620	3	0,8073	0,9164	0,8375	0,7149	0,7714
Ward	3	40	1620	3	0,8259	0,9091	0,8414	0,6979	0,7629
K-Means	3	40	1690	2	0,7257	0,6032	0,1304	0,2010	0,1582
Complete	3	40	1690	2	0,7257	0,6032	0,1304	0,2010	0,1582
Ward	3	40	1690	2	0,7257	0,6032	0,1304	0,2010	0,1582
K-Means	3	40	1750	2	0,9744	1,0000	1,0000	1,0000	1,0000
Complete	3	40	1750	2	0,9744	1,0000	1,0000	1,0000	1,0000
Ward	3	40	1750	2	0,9744	1,0000	1,0000	1,0000	1,0000
K-Means	3	40	1760	2	0,8640	1,0000	1,0000	1,0000	1,0000
Complete	3	40	1760	2	0,8640	1,0000	1,0000	1,0000	1,0000
Ward	3	40	1760	2	0,8640	1,0000	1,0000	1,0000	1,0000
K-Means	3	40	1800	3	0,7013	0,8229	0,9405	0,6448	0,7651
Complete	3	40	1800	3	0,8233	1,0000	1,0000	1,0000	1,0000
Ward	3	40	1800	3	0,7013	0,8229	0,9405	0,6448	0,7651
K-Means	3	50	4	3	0,8907	0,7336	0,0886	0,0599	0,0715
Complete	3	50	4	3	0,8822	0,7343	0,0929	0,0632	0,0752
Ward	3	50	4	3	0,8901	0,7337	0,0877	0,0594	0,0709
K-Means	3	50	20	2	0,8527	0,7004	0,2234	0,3594	0,2755
Complete	3	50	20	2	0,8107	0,7498	0,3929	0,4863	0,4346
Ward	3	50	20	2	0,8527	0,7004	0,2234	0,3594	0,2755
K-Means	3	50	30	6	0,7239	0,6995	0,7455	0,7182	0,7316
Complete	3	50	30	6	0,7002	0,6085	0,6556	0,6996	0,6768
Ward	3	50	30	6	0,7061	0,6405	0,6721	0,7254	0,6978
K-Means	3	50	50	4	0,6059	0,5242	0,3828	0,1348	0,1993
Complete	3	50	50	4	0,6058	0,5409	0,1247	0,0495	0,0708
Ward	3	50	50	4	0,5952	0,5161	0,3026	0,1089	0,1601
K-Means	3	55	50	3	0,7107	0,6231	0,4722	0,2226	0,3025
Complete	3	55	50	3	0,7076	0,6410	0,5102	0,2433	0,3295
Ward	3	55	50	3	0,7076	0,6410	0,5102	0,2433	0,3295
K-Means	3	60	20	4	0,5702	0,5696	0,5048	0,0249	0,0474

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Complete	3	60	20	4	0,6211	0,7190	0,6521	0,0496	0,0922
Ward	3	60	20	4	0,5848	0,5745	0,4866	0,0242	0,0460
K-Means	3	60	21	2	0,7337	0,8711	0,3642	0,0598	0,1027
Complete	3	60	21	2	0,7451	0,8605	0,0274	0,0048	0,0081
Ward	3	60	21	2	0,7451	0,8605	0,0274	0,0048	0,0081
K-Means	3	60	40	2	0,6338	0,6900	0,1257	0,0180	0,0315
Complete	3	60	40	2	0,4755	0,8541	0,3706	0,0799	0,1314
Ward	3	60	40	2	0,6200	0,6979	0,1072	0,0157	0,0274
K-Means	3	60	60	3	0,7487	0,8532	0,5253	0,3021	0,3836
Complete	3	60	60	3	0,7742	0,9197	0,4302	0,5074	0,4656
Ward	3	60	60	3	0,7487	0,8532	0,5253	0,3021	0,3836
K-Means	3	75	10	2	0,7337	0,6608	0,1121	0,0686	0,0851
Complete	3	75	10	2	0,7337	0,6608	0,1121	0,0686	0,0851
Ward	3	75	10	2	0,7337	0,6608	0,1121	0,0686	0,0851
K-Means	3	90	70	13	0,5813	0,3922	0,5239	0,1247	0,2015
Complete	3	90	70	13	0,6414	0,5877	0,5485	0,1755	0,2659
Ward	3	90	70	13	0,5522	0,3694	0,5200	0,1212	0,1966
K-Means	3	90	79	2	0,9351	1,0000	1,0000	1,0000	1,0000
Complete	3	90	79	2	0,9351	1,0000	1,0000	1,0000	1,0000
Ward	3	90	79	2	0,9351	1,0000	1,0000	1,0000	1,0000
K-Means	3	90	190	2	0,8361	0,7148	0,1154	0,0019	0,0037
Complete	3	90	190	2	0,7519	0,9750	0,0032	0,0003	0,0006
Ward	3	90	190	2	0,8361	0,7148	0,1154	0,0019	0,0037
K-Means	3	90	226	2	0,8745	0,8909	0,0680	0,0046	0,0086
Complete	3	90	226	2	0,8734	0,8925	0,3259	0,0217	0,0407
Ward	3	90	226	2	0,8752	0,8910	0,0180	0,0012	0,0023
K-Means	3	90	320	2	0,7187	0,7541	0,3886	0,4886	0,4329
Complete	3	90	320	2	0,6139	0,9742	0,9046	0,9124	0,9085
Ward	3	90	320	2	0,7162	0,7259	0,3046	0,4258	0,3551
K-Means	3	90	470	2	0,7384	0,6214	0,1370	0,1198	0,1278
Complete	3	90	470	2	0,8737	0,7682	0,0194	0,2216	0,0358
Ward	3	90	470	2	0,7260	0,6637	0,3068	0,2655	0,2847
K-Means	3	90	550	5	0,6311	0,5858	0,3925	0,0198	0,0378
Complete	3	90	550	5	0,3687	0,6978	0,3694	0,0243	0,0455
Ward	3	90	550	5	0,6018	0,5749	0,4143	0,0209	0,0398
K-Means	3	90	920	2	0,8020	0,7133	0,2523	0,0648	0,1032
Complete	3	90	920	2	0,7338	1,0000	1,0000	1,0000	1,0000
Ward	3	90	920	2	0,8020	0,7133	0,2523	0,0648	0,1032
K-Means	3	90	930	4	0,7886	0,6374	0,4726	0,0471	0,0856
Complete	3	90	930	4	0,7885	0,6401	0,4730	0,0473	0,0860
Ward	3	90	930	4	0,7885	0,6401	0,4730	0,0473	0,0860

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
K-Means	3	90	940	4	0,6986	0,5542	0,4311	0,1935	0,2671
Complete	3	90	940	4	0,7354	0,6021	0,2891	0,1778	0,2202
Ward	3	90	940	4	0,6996	0,5452	0,4228	0,1889	0,2611
K-Means	3	90	950	2	0,4628	0,6553	0,0453	0,0192	0,0269
Complete	3	90	950	2	0,4370	0,6501	0,0469	0,0205	0,0285
Ward	3	90	950	2	0,4628	0,6553	0,0453	0,0192	0,0269
K-Means	3	90	1010	2	0,7309	0,6320	0,1890	0,0997	0,1306
Complete	3	90	1010	2	0,7309	0,6320	0,1890	0,0997	0,1306
Ward	3	90	1010	2	0,7309	0,6320	0,1890	0,0997	0,1306
K-Means	3	90	1150	7	0,7465	0,5643	0,6448	0,6258	0,6352
Complete	3	90	1150	7	0,7465	0,5643	0,6448	0,6258	0,6352
Ward	3	90	1150	7	0,7465	0,5643	0,6448	0,6258	0,6352
K-Means	3	90	1170	2	0,8107	0,7683	0,0283	0,0150	0,0196
Complete	3	90	1170	2	0,8096	0,7670	0,0564	0,0302	0,0393
Ward	3	90	1170	2	0,8096	0,7670	0,0564	0,0302	0,0393
K-Means	3	90	1210	2	0,9237	0,7758	0,0555	0,0215	0,0310
Complete	3	90	1210	2	0,9237	0,7758	0,0555	0,0215	0,0310
Ward	3	90	1210	2	0,9237	0,7758	0,0555	0,0215	0,0310
K-Means	3	90	1280	2	0,9220	0,8030	0,4892	0,5755	0,5289
Complete	3	90	1280	2	0,9220	0,8030	0,4892	0,5755	0,5289
Ward	3	90	1280	2	0,9220	0,8030	0,4892	0,5755	0,5289
K-Means	3	90	1480	2	0,6017	0,8206	0,5532	0,5580	0,5556
Complete	3	90	1480	2	0,6017	0,8206	0,5532	0,5580	0,5556
Ward	3	90	1480	2	0,6017	0,8206	0,5532	0,5580	0,5556
K-Means	3	90	1570	2	0,8179	0,7035	0,5440	0,4766	0,5081
Complete	3	90	1570	2	0,8179	0,7035	0,5440	0,4766	0,5081
Ward	3	90	1570	2	0,8179	0,7035	0,5440	0,4766	0,5081
K-Means	3	90	1640	2	0,8926	1,0000	1,0000	1,0000	1,0000
Complete	3	90	1640	2	0,8926	1,0000	1,0000	1,0000	1,0000
Ward	3	90	1640	2	0,8926	1,0000	1,0000	1,0000	1,0000
K-Means	3	90	1720	3	0,6399	0,7005	0,0392	0,0170	0,0237
Complete	3	90	1720	3	0,9046	0,8434	0,0218	0,0218	0,0218
Ward	3	90	1720	3	0,6327	0,6307	0,2288	0,0769	0,1151
K-Means	3	90	1820	2	0,9320	1,0000	1,0000	1,0000	1,0000
Complete	3	90	1820	2	0,9320	1,0000	1,0000	1,0000	1,0000
Ward	3	90	1820	2	0,9320	1,0000	1,0000	1,0000	1,0000
K-Means	4	10	40	3	0,5622	0,4349	0,0881	0,0670	0,0761
Complete	4	10	40	3	0,5551	0,4228	0,0584	0,0449	0,0508
Ward	4	10	40	3	0,5562	0,4241	0,0604	0,0465	0,0526
K-Means	4	10	81	2	0,7924	0,8335	0,3916	0,5448	0,4556
Complete	4	10	81	2	0,8752	0,7688	0,0253	0,2354	0,0457

Próxima página

Tabela A.1 – *Continuação*

Método	C	A	T	Grupos	S	F	H	Co.	V
Ward	4	10	81	2	0,7855	0,8441	0,4295	0,5714	0,4904
K-Means	4	10	110	2	0,9010	0,7445	0,0356	0,0658	0,0462
Complete	4	10	110	2	0,9010	0,7445	0,0356	0,0658	0,0462
Ward	4	10	110	2	0,9010	0,7445	0,0356	0,0658	0,0462
K-Means	4	10	220	2	0,7264	0,6715	0,1853	0,0536	0,0831
Complete	4	10	220	2	0,7028	0,6669	0,1593	0,0463	0,0718
Ward	4	10	220	2	0,7028	0,6669	0,1593	0,0463	0,0718
K-Means	4	10	260	3	0,5669	0,6228	0,3284	0,3136	0,3208
Complete	4	10	260	3	0,5363	0,6241	0,2812	0,2860	0,2836
Ward	4	10	260	3	0,5669	0,6228	0,3284	0,3136	0,3208
K-Means	4	10	280	3	0,5877	0,6913	0,4616	0,3151	0,3745
Complete	4	10	280	3	0,4065	0,7144	0,3503	0,2892	0,3168
Ward	4	10	280	3	0,5877	0,6913	0,4616	0,3151	0,3745
K-Means	4	10	860	2	0,9800	0,7805	0,0058	0,0509	0,0104
Complete	4	10	860	2	0,9800	0,7805	0,0058	0,0509	0,0104
Ward	4	10	860	2	0,9800	0,7805	0,0058	0,0509	0,0104

Apêndice B

Artigo Publicado

Cleiton Monteiro, Vinıcio Mendes, Giovanni Comarela, and Sabrina A. Silveira. **Using supervised learning successful descriptors to perform protein structural classification through unsupervised learning**. In IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2018), pp. 75-78. IEE, 2018.