Bruno Almeida Pimentel

MÉTODOS DE AGRUPAMENTO DIFUSO MULTIVARIADO BASEADOS NO *FUZZY C-MEANS*

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE
2017

Bruno Almeida Pimentel

MÉTODOS DE AGRUPAMENTO DIFUSO MULTIVARIADO BASEADOS NO *FUZZY C-MEANS*

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Doutor em Ciência da Computação.

Orientadora: *Profa. Dra. Renata Maria Cardoso Rodrigues de Souza*

RECIFE
2017

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

P644m Pimentel, Bruno Almeida
Métodos de agrupamento difuso multivariado baseados no *Fuzzy C-means* / Bruno Almeida Pimentel. – 2017.
143 f.:il., fig., tab.

Orientadora: Renata Maria Cardoso Rodrigues de Souza.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2017.
Inclui referências e apêndices.

1. Inteligência computacional. 2. Análise de agrupamento. I. Souza, Renata Maria Cardoso Rodrigues de (orientadora). II. Título.

006.3 CDD (23. ed.) UFPE- MEI 2017-226

Bruno Almeida Pimentel

Métodos de Agrupamento Difuso Multivariado Baseados no *Fuzzy C-Means*

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de **Doutor** em Ciência da Computação.

Aprovado em: 21/09/2017

Orientadora: Profa. Dra. Renata Maria Cardoso Rodrigues de Souza

BANCA EXAMINADORA

Prof. Dr. Adriano Lorena Inácio de Oliveira
Centro de Informatica / UFPE

Prof. Dr. George Darmiton da Cunha Cavalcanti
Centro de Informatica / UFPE

Prof. Dr. Cleber Zanchettin
Centro de Informatica / UFPE

Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho
Instituto de Ciências Matemáticas e de Computação / USP

Profa. Dra. Roberta Andrade de Araujo Fagundes
Escola Politécnica de Pernambuco / UPE

Dedico esse trabalho aos meus pais, por terem ensinado o verdadeiro valor dos estudos.

Agradecimentos

Agradeço à minha família por ter me ajudado a chegar nesta etapa da minha vida, com incentivo e compreensão, em especial aos meus pais que estão me acompanhando e orientando em todos os momentos. Por terem me ensinado o valor do estudo e que a busca pelo conhecimento e pela verdadeira vocação devem ser características essenciais presentes em qualquer cidadão.

À professora, orientadora e amiga Renata de Souza, que desde os primeiros semestres do curso de Ciência da Computação me auxiliou nessa construção do conhecimento, permitindo, primeiramente, o ingresso na monitoria de Estatística, a qual deu ocasião a ajudar outros colegas e aprender mais sobre a disciplina. Em seguida, a participação de pesquisas através da Iniciação Científica, atividade que me fez ficar ainda mais admirado pela ciência. Guiou-me no programa de Mestrado e mais recentemente orientou-me na pesquisa e desenvolvimento deste trabalho.

Aos membros da banca examinadora por aceitaram participar e pelas contribuições e direcionamentos que vieram no intuito de enriquecer este trabalho.

Agradeço também aos meus colegas de turma, com os quais debatemos ideias e discutimos sobre os mais diversos temas ajudando na minha formação acadêmica e profissional. Aos amigos de pesquisa, que antes mesmo do Doutorado estamos adquirindo conhecimentos juntos.

A todos que participaram direta ou indiretamente contribuindo para a realização deste trabalho. Finalmente, ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo auxílio financeiro.

Ninguém ignora tudo. Ninguém sabe tudo. Todos nós sabemos alguma coisa.
Todos nós ignoramos alguma coisa. Por isso aprendemos sempre.

—PAULO FREIRE

Resumo

A prática de agrupar objetos de acordo com as similaridades e propriedades observadas é uma atividade importante para muitos ramos da ciência. A sua importância deve-se ao fato que a organização dos dados em grupos é uma forma fundamental para entender e aprender sobre eles. Em Biologia, por exemplo, existe a preocupação de dividir os diferentes animais ou plantas em grupos para melhor entendimento das funções biológicas. Em muitas problemas, além de informar a qual grupo um determinado objeto pertence, é necessário entender quão similar este objeto está para todos os grupos da partição devido à imprecisão ou incerteza dos dados, surgindo, assim, o agrupamento difuso. O principal método de agrupamento difuso bastante conhecido é o *Fuzzy C-Means* (FCM), o qual tem algumas desvantagens tal como considerar que todos os grupos possuem formas esféricas. Outra desvantagem é que não existe a possibilidade de analisar qual variável (ou um sub-conjunto delas) foi mais importante para definir o valor final do grau de pertinência. Este trabalho apresenta diferentes métodos de agrupamento usando a abordagem difusa presentes na literatura atual e introduz métodos de agrupamento difuso onde os graus de pertinência são multivariados. Desta forma, dado um objeto, é possível calcular o grau dele pertencer a um dado grupo segundo uma variável. A partir deste tipo de grau de pertinência multivariado, duas vantagens podem ser apontadas: 1 - capacidade de interpretar a pertinência de cada objeto para um dado grupo segundo cada variável; 2 - obtenção de mais informação dos dados levando a uma maior qualidade de agrupamento. O objetivo deste trabalho é propor duas categorias de métodos: a primeira é baseada no *Fuzzy C-Means* e a segunda é baseada no *Possibilistic Fuzzy C-Means*. Além disso, também são propostos índices de interpretação para avaliar a qualidade do agrupamento para grupo e variável a partir da partição difusa obtida por cada método proposto. Com o objetivo de avaliar o desempenho dos métodos, um estudo comparativo em relação ao agrupamento difuso usando o experimento Monte Carlo é realizado. Foram planejados experimentos com dados sintéticos e reais e um índice de validação é usado para avaliar os métodos. Além disso, aplicação com dados biológicos é apresentada mostrando a utilidade dos métodos propostos. Os resultados mostraram que os métodos multivariados são preferíveis quando as variáveis são independentes e possuem variabilidades intra-classe diferentes.

Palavras-chave: Agrupamento Difuso. Grau Multivariado. *Fuzzy C-Means*. *Possibilistic Fuzzy C-Means*. Aplicação.

Abstract

The practice of classifying objects according to the observed similarities and properties is an important activity for many branches of science. Its importance is due to the fact that the organization of data into groups is a fundamental mode to understand and learn about ones. In Biology, for example, there is concern divide the different animals or plants into groups for better understanding of biological functions. In many problems, besides informing the group which a particular object belongs, it is necessary to understand how this object is similar for all groups due to of the vagueness or uncertainty of the data, emerging, so the fuzzy clustering. The primary method of fuzzy clustering is the Fuzzy C-Means (FCM), which has some disadvantages as considering that all groups have spherical shapes. Another disadvantage is that there is not the possibility to analyze which variable (or a subset of them) was more important to set the final value of the degree of membership. This work presents different clustering methods using fuzzy approach present in the current literature and introduces fuzzy clustering methods where the degrees of membership are multivariate. Thus, given an object, it is possible to calculate the degree it belongs to a group according to a given variable. From this type of degree multivariate relevance, two advantages can be pointed out: 1 - ability to interpret the relevance of each object for a given group according to each variable; 2 - getting more information from the data leading to a better quality of clustering. The objective of this work is to propose two types of methods: the first one is based on the Fuzzy C-Means and the second one is based on the Possibilistic Fuzzy C-Means. Moreover, interpretation indices are also proposed for assessing the quality of the clustering according to each cluster and variable from a fuzzy partition obtained by each proposed method. Aiming to evaluate the performance of the methods, a comparative study with respect to fuzzy clustering using the Monte Carlo experiment is carried out. Experiments with synthetic and real data and a validation index is used to evaluate the methods were planned. Furthermore, application with biological data is presented showing the usefulness of the proposed methods. The results showed that multivariate methods are preferable when the variables are independent and have different intra-class variabilities.

Keywords: Fuzzy Clustering. Multivariate Degree. Fuzzy C-Means. Possibilistic Fuzzy C-Means. Application.

Lista de Figuras

1.1 Conjunto de dados simples. 19
1.2 Resultados obtidos pelo FCM (esquerda) e sua versão multivariada (direita). . . 20
1.3 Conjunto de dados Coluna Vertebral de acordo com as variáveis "Raio pélvico" e "Grau de espondilolistese". 21
1.4 Análise de Coluna Vertebral com suas variáveis separadas: "Raio pélvico" em cima e "Grau de espondilolistese" embaixo. 21

2.1 Variação do grau de pertinência em função do somatótio para diferentes valores de m. 28

3.1 Dois objetos equidistantes do protótipo. 43
3.2 Dois protótipos e um objeto em uma região de ambiguidade. 47
3.3 Partições contendo grupos de diferentes formas e tamanhos. 50
3.4 Conjunto contendo dados aberrantes. 54
3.5 Conjunto contendo dados aberrantes e grupos de diferentes formas e tamanhos. 59

4.1 Conjunto de dados 1 (esquerda) e 2 (direita). 76
4.2 Conjunto de dados 3 (esquerda) e 4 (direita). 76
4.3 Conjunto de dados 5 (esquerda) e 6 (direita). 78
4.4 Conjunto de dados 7. 79
4.5 Conjunto de dados 8 (esquerda) e 9 (direita). 81
4.6 Conjunto de dados 10 (esquerda) e 11 (direita). 81
4.7 Conjunto de dados 12 com 0% de ruido (esquerda) e 30% (direita). 83
4.8 Conjunto de dados 13 com 0% de ruido (esquerda) e 30% (direita). 84

5.1 Relação entre célula, cromossomo, DNA e gene. 95
5.2 Exemplo de uma estrutura de DNA com suas ligações entre bases. 95
5.3 Constituição de um *microarray* a partir de amostras de DNA e utilização de *lasers* para a excitação dos rótulos fluorescentes. 96
5.4 Dispersão do conjunto de dados sintético. 99

Lista de Tabelas

4.1 Parâmetros para os conjuntos de dados 1 e 2. 75
4.2 Parâmetros para os conjuntos de dados 3 e 4. 75
4.3 Média (maiores valores em negrito) e desvio padrão (em parênteses) do índice FR para os métodos de agrupamento e conjunto de dados 1-4. 76
4.4 Parâmetros para os conjuntos de dados 5 e 6. 77
4.5 Parâmetros para o conjunto de dados 7. 78
4.6 Média e desvio padrão (em parênteses) do índice FR para os métodos de agrupamento e conjuntos de dados 5-7. 79
4.7 Parâmetros para os conjuntos com dados 8 e 9. 80
4.8 Parâmetros para os conjuntos com dados 10 e 11. 80
4.9 Média e desvio padrão (entre parênteses) do índice FR (maiores valores em negrito) para conjunto de dados 8, 9 e 10. 80
4.10 Média e desvio padrão (entre parênteses) do índice FR (maiores valores em negrito) para conjunto de dados 11. 82
4.11 Parâmetros para os conjuntos com dados 12 e 13. 83
4.12 Média e desvio padrão (entre parênteses) do índice FR (maiores valores em negrito) para conjunto com dados ruidosos e classes esféricas. 84
4.13 Média e desvio padrão (entre parênteses) do índice FR (maiores valores em negrito) para conjunto com dados ruidosos e classes elípticas. 85
4.14 Descrição de conjuntos de dados reais da UCI. 86
4.15 Média e desvio padrão (entre parênteses) do índice FR (maiores valores em negrito) para métodos de agrupamento e conjuntos de dados da UCI. 87
4.16 Ordenação dos métodos de agrupamento quanto às maiores médias do índice FR para os conjuntos de dados reais da UCI. 88
4.17 Descrição das variáveis para o conjunto de dados Abalone. 89
4.18 Índice de heterogeneidade global para os métodos multivariados. 89
4.19 Índice de heterogeneidade global e inter-grupo de acordo com cada variável e método. 90
4.20 Contribuição relativa de grupos de medidas baseadas na soma de quadrados para os métodos MFCM, MFCM-G e MFCM-D. 90

4.21 Contribuição relativa de grupos de medidas baseadas na soma de quadrados para os métodos MPFCM e MPFCM-D. 90
4.22 Índices de interpretação por grupo e variável para o método MFCM. 91
4.23 Índices de interpretação por grupo e variável para o método MFCM-G. 91
4.24 Índices de interpretação por grupo e variável para o método MFCM-D. 92
4.25 Índices de interpretação por grupo e variável para o método MPFCM. 92
4.26 Índices de interpretação por grupo e variável para o método MPFCM-D. 93

5.1 Parâmetros para o conjunto de dados sintético. 99
5.2 Média e desvio padrão (em parênteses) de a_{ij} dos métodos multivariados e conjunto de dados sintético. 100
5.3 Descrição de conjuntos de dados de expressão de genes. 101
5.4 Média e desvio padrão (entre parênteses) do índice de FR (maiores valores em negrito) para os métodos de agrupamento e todos os conjuntos de dados. 102
5.5 Valores selecionados de a_{ij} e correspondentes genes. 103

B.1 Métodos e medianas. 135
B.2 Teste de Friedman para configurações 1-10. 136
B.3 Teste de Friedman para Configuração 11: variações no ângulo. 136
B.4 Teste de Friedman para Configuração 12: variações no percentual de ruído. . . 136
B.5 Teste de Friedman para Configuração 13: variações no percentual de ruído. . . 136
B.6 Teste de Friedman para conjuntos de dados da UCI — parte 1. 136
B.7 Teste de Friedman para conjuntos de dados da UCI — parte 2. 136
B.8 Teste de Friedman para conjuntos de genes. 136
B.9 Métodos multivariados e suas suposições a respeito da configuração dos conjuntos de dados. 137
B.10 Testes de Wilcoxon-Mann-Whitney para MFCM: Configuração 3. 137
B.11 Teste de Wilcoxon-Mann-Whitney para MFCM-G: Configuração 1. 138
B.12 Teste de Wilcoxon-Mann-Whitney para MFCM-D: Configuração 2. 138
B.13 Teste de Wilcoxon-Mann-Whitney para MPFCM: Configuração 12 (30%). . . . 138
B.14 Teste de Wilcoxon-Mann-Whitney para MPFCM-D: Configuração 13 (30%). . 138
B.15 Testes de Wilcoxon-Mann-Whitney para MFCM: dados reais da UCI. 139
B.16 Testes de Wilcoxon-Mann-Whitney para MFCM-G: dados reais da UCI. 140
B.17 Testes de Wilcoxon-Mann-Whitney para MFCM-D: dados reais da UCI. 140
B.18 Testes de Wilcoxon-Mann-Whitney para MPFCM: dados reais da UCI. 141
B.19 Testes de Wilcoxon-Mann-Whitney para MPFCM-D: dados reais da UCI. . . . 141

B.20 Testes de Wilcoxon-Mann-Whitney para MFCM: dados de gene. 142
B.21 Testes de Wilcoxon-Mann-Whitney para MFCM-G: dados de gene. 142
B.22 Testes de Wilcoxon-Mann-Whitney para MFCM-D: dados de gene. 143
B.23 Testes de Wilcoxon-Mann-Whitney para MPFCM: dados de gene. 143
B.24 Testes de Wilcoxon-Mann-Whitney para MPFCM-D: dados de gene. 143

Lista de Acrônimos

FCM	Método *Fuzzy C-Means*
FCM-C[1]	Método *Fuzzy C-Means* com Ponderação por Variável
FCM-C[2]	Método *Fuzzy C-Means* com Distância Adaptativa por Classe
PFCM	Método *Possibilistic Fuzzy C-Means*
W-FCM	Método *Fuzzy C-Means* Ponderado
MFCM	Método *Fuzzy C-Means* Multivariado
MFCM-G	Método *Fuzzy C-Means* Multivariado com Ponderação no Grau
MFCM-D	Método *Fuzzy C-Means* Multivariado com Ponderação na Distância
MPFCM	Método *Possibilistic Fuzzy C-Means* Multivariado
MPFCM-D	Método *Possibilistic Fuzzy C-Means* Multivariado com Ponderação na Distância
FR	Índice de validade de agrupamento difuso
UCI	Universidade da Califórnia Irvine

Sumário

1 INTRODUÇÃO 17
1.1 **Motivação** . . . 17
1.2 **Objetivos** . . . 22
1.3 **Organização da Tese** . . . 23

2 MÉTODOS DE AGRUPAMENTO DIFUSO 25
2.1 **Introdução** . . . 25
2.2 **Fuzzy C-Means** . . . 26
2.3 **Fuzzy C-Means com Ponderação por Variável** . . . 29
2.4 **Fuzzy C-Means com Distância Adaptativa** . . . 30
2.5 **Fuzzy C-Means Robusto** . . . 32
2.6 **Possibilistic Fuzzy C-Means** . . . 35
2.7 **Hierarchical Fuzzy Clustering** . . . 36
2.8 **Fuzzy C-Means Ponderado** . . . 39

3 MÉTODOS DE AGRUPAMENTO DIFUSO MULTIVARIADO 41
3.1 **Introdução** . . . 41
3.2 **Métodos Propostos** . . . 42
3.2.1 Método Fuzzy C-Means Multivariado . . . 42
3.2.2 Método Fuzzy C-Means Multivariado com Ponderação no Grau de Pertinência . . . 46
3.2.3 Método Fuzzy C-Means Multivariado com Ponderação na Distância . . . 50
3.2.4 Método Possibilistic Fuzzy C-Means Multivariado . . . 53
3.2.5 Método Possibilistic Fuzzy C-Means Multivariado Ponderado . . . 58
3.3 **Interpretação de Partição e Grupo** . . . 62
3.3.1 Medidas baseadas em Soma de Quadrados . . . 62
3.3.2 Índices de Interpretação . . . 69
3.3.2.1 *Índices de Interpretação de Partição* . . . 69
3.3.2.2 *Índices de Interpretação de Grupos* . . . 70

4 EXPERIMENTOS E RESULTADOS 72
4.1 **Introdução** . . . 72

4.2 **Dados Sintéticos** 74
4.2.1 Formato único das classes 74
4.2.2 Mistura de Formas 77
4.2.3 Variáveis Correlacionadas 79
4.2.4 Dados Ruidosos 82
4.3 **Conjunto de Dados Reais da UCI** 85
4.4 **Interpretação de Partições e Grupos Difusos: Conjunto de Dados Abalone** 88

5 APLICAÇÃO EM DADOS BIOLÓGICOS 94
5.1 **Introdução a Bioinformática** 94
5.2 **Motivação** 97
5.3 **Objetivo** 98
5.4 **Uma Estratégia para Interpretar Graus de Pertinência Multivariados** . . . 98
5.5 **Aplicação** 101

6 CONCLUSÕES 105
6.1 **Conclusão** 105
6.2 **Publicações** 107
6.3 **Trabalhos Futuros** 109

REFERÊNCIAS 110

GLOSSÁRIO 120

APÊNDICE A - PROVAS 124
A.1 **Prova da Proposição 3.2.1** 124
A.2 **Prova da Proposição 3.2.2** 125
A.3 **Prova da Proposição 3.2.3** 126
A.4 **Prova da Proposição 3.2.4** 127
A.5 **Prova da Proposição 3.2.7** 127
A.6 **Prova da Proposição 3.2.8** 128
A.7 **Prova da Proposição 3.2.9** 129
A.8 **Prova da Proposição 3.2.12** 130
A.9 **Prova da Proposição 3.2.17** 131
A.10 **Prova da Proposição 3.2.23** 132

A.11 **Prova da Proposição 3.3.2** 133

APÊNDICE B - TESTES **135**

B.1 **Testes de Friedman** 135

B.2 **Testes de Wilcoxon-Mann-Whitney** 137

B.2.1 Dados Sintéticos 137

B.2.2 Dados Reais da UCI 139

B.2.3 Dados de Genes 142

1 INTRODUÇÃO

Aprender é a única de que a mente nunca se cansa, nunca tem medo e nunca se arrepende.

—LEONARDO DA VINCI

1.1 Motivação

Aprendizado de Máquina é um importante ramo da Inteligência Artificial, cujo principal objetivo é estudar e propor métodos capazes de aprender a partir de dados (Anderson, 1986; Witten et al., 2011). Em muitos problemas, Aprendizado de Máquina pode ser aplicada quando o processo de extração de informações a partir de dados é complicado (Anzai, 1992). Com base no tipo de entrada disponível durante o treinamento, os algoritmos de aprendizado de máquina podem ser supervisionados ou não-supervisionados (MacKay, 2003; Pedregosa et al., 2011). No primeiro, os algoritmos são treinados a partir de dados rotulados, enquanto no segundo algoritmos são treinados a partir de dados não rotulados. Para muitos problemas, não é possível obter rótulos de dados ou o custo computacional para a obtenção é elevado (Blum e Chawla, 2001; Breve et al., 2012), portanto, abordagens para aprendizado não-supervisionada como agrupamento são mais adequados para resolver estes problemas. O principal objetivo dos algoritmos de agrupamento é reunir um conjunto de objetos tais que os objetos no mesmo grupo são mais semelhantes do que em grupos diferentes. Métodos de agrupamento tem como meta otimizar a função objetivo adequada e geralmente convergem para uma partição ótima local (Babu e Murty, 1994; Bai et al., 2013). Com o crescente interesse em compreender automaticamente, processar e sintetizar dados, diversos domínios de aplicação, tais como reconhecimento de padrões, aprendizado de máquina, mineração de dados, visão computacional

e biologia computacional usam algoritmos de agrupamento (Duda et al., 2001; Jain, 1999; Xu e Wunsch, 2005; Gan et al., 2007).

De acordo com a forma como algoritmos de agrupamento encontram grupos de dados, eles podem ser classificados em: hierárquico e particional (Jain, 1999; Jain e Dubes, 1988; Kuo et al., 2013). Na primeira categoria, os algoritmos produzem uma série encadeada de partições com base em um critério aglomerativo ou divisivo. Se o algoritmo hierárquico é aglomerativo, começa-se com cada objeto em um grupo e sucessivamente junta grupos com base nas semelhanças. Se o algoritmo hierárquico é divisivo, começa-se com todos os objetos em um único grupo e divisões são executadas até encontrar grupos com um único objeto. A distância entre dois grupos pode ser medida utilizando critério de ligação que pode mudar a maneira como os algoritmos hierárquicos encontram aglomerados (Szekely e Rizzo, 2005; Krishnamurthy et al., 2012). Por outro lado, a segunda categoria de algoritmos de agrupamento, o particional, obtem uma única partição de dados em vez de uma série encadeada de partições. A escolha do número de grupos desejados pode ser um problema, já que nem sempre há *a priori* informações suficientes do conjunto de dados para definir este número (Milligan e Cooper, 1985). Existem vários trabalhos sobre como encontrar automaticamente o número de grupos (Fraley e Raftery, 1998; Hruschka et al., 2009). Entretanto, os métodos particionais têm vantagens em muitos conjuntos de dados, uma vez que a complexidade algorítmica é menor do que a dos métodos hierárquicos (Jain, 1999; Xu e Wunsch, 2005).

Métodos particionais podem ser classificados em duas principais categorias: *hard* e *soft* (Jain, 1999). Na primeira categoria, objetos pertencem a exatamente um grupo, onde o grau de pertinência de um objeto para um grupo pode estar no conjunto $\{0,1\}$: 1 se o objeto pertencer ao grupo ou 0 caso contrário (Pal et al., 2005). Um dos métodos particionais *hard* mais populares é o *K-Means* (Jain, 2010) proposto por MacQueen (1967). Na segunda categoria de métodos particionais, os objetos têm graus de pertinência para todos os grupos, onde podem estar no intervalo $[0,1]$. Métodos *soft* podem ser difusos, probabilísticos ou possibilísticos (Pal et al., 2005). A principal vantagem dos métodos *soft* é sua capacidade de encontrar grupos sobrepostos, isto é, onde existem objetos de diferentes grupos com alta similaridade. Além disso, uma vez que os graus de pertinência não são discretos, os métodos do tipo *soft* dão a possibilidade de interpretar o valor do grau de pertinência: quão semelhante um elemento é de um dado grupo. O método de agrupamento *Fuzzy C-Means* (FCM) proposto por Bezdek (1981) é amplamente conhecido e poderoso na abordagem difusa (Yang, 1993). A sua principal desvantagem é o seu baixo desempenho em conjuntos com presença de dados aberrantes (*outliers* ou ruidos). O principal representante da abordagem possibilística é o método de agrupamento *Possibilistic C-Means* (PCM) proposto por Krishnapuram e Keller (1993) para suprir a deficiência do FCM.

Quanto à abordagem difusa, muitas extensões do método FCM foram propostas na literatura. Zhang e Chen (2003) propuseram uma nova versão do FCM e PCM substituindo a distância euclidiana original por uma distância baseada em *kernel*. Pal et al. (1997) apresentaram um novo modelo denominado *Fuzzy-Possibilistic C-Means* (FPCM) que mistura propriedades dos métodos FCM e PCM. Pal et al. (2005) propôs uma hibridização de PCM e FCM (abreviado por PFCM), que evita diversos problemas de PCM, FCM e FPCM tais como protótipos coincidentes, baixa robustez a dados ruidosos e escalabilidade. Keller e Klawonn (2000) introduziram uma função objetivo que atribui um parâmetro de influência para cada variável de dados para cada grupo. De Carvalho et al. (2006a) apresentou um método de agrupamento difuso com base em distâncias quadráticas adaptativas onde pesos podem ser iguais para todos os grupos ou diferentes de um grupo para outro. Ferreira e De Carvalho (2014) introduziram um método difuso com *kernel* por variável. Embora estes métodos possam produzir melhores resultados do que o FCM tradicional, eles não consideram que as diferentes variáveis podem produzir diferentes graus de pertinência. Portanto, o objetivo deste trabalho é propor métodos difusos cujos graus de pertinência de um dado objeto sejam, também, por variável. Desta forma, os graus de pertinência calculados nos métodos de agrupamento difuso tornam-se multivariados (por simplicidade, estes métodos serão chamados multivariados).

Para compreender a motivação dos métodos multivariados, considere dois exemplos. O primeiro é um exemplo simples (*toy*) formado por duas classes onde a Variável 1 possui uma variabilidade maior que a Variável 2 para cada classe. Além disso, através da Variável 2, é possível diferenciar bem as duas classes, ao contrário da Variável 1. A Figura 1.1 mostra como essas classes estão organizadas.

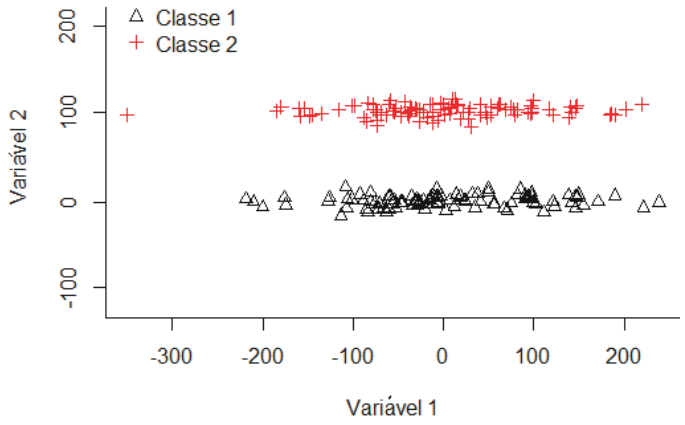

Figura 1.1: Conjunto de dados simples.

Se o método FCM fosse aplicado a esse conjunto de dados, a partição final obtida seria como apresentada na Figura 1.2 (esquerda). Enquanto se a versão multivariada desse método fosse aplicada a esse conjunto, obteria-se a partição como apresentada na Figura 1.2 (direita). Nota-se que a partição proveniente da versão multivariada está mais próxima da original. Isso acontece porque a Variável 2 é a que mais distingue os grupos e essa informação é levada em conta utilizando graus multivariados.

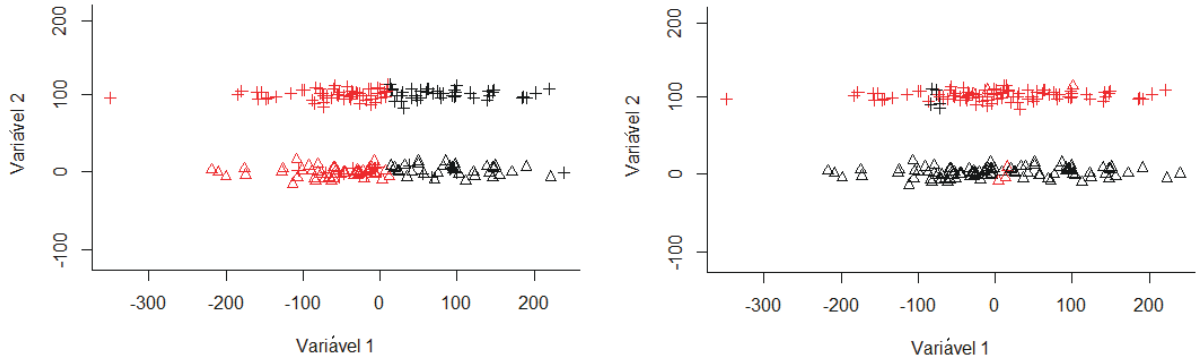

Figura 1.2: Resultados obtidos pelo FCM (esquerda) e sua versão multivariada (direita).

Para o segundo exemplo, considere o conjunto de dados Coluna Vertebral. Trata-se de informações biomédicas e pode ser obtido no repositório de aprendizado de máquina da UCI (Bache e Lichman, 2013). Os dados foram coletados durante uma residência média e diz respeito a 6 características biomecânicas da coluna vertebral de pacientes: incidência pélvica, inclinação pélvica, ângulo da lordose lombar, inclinação sacral, raio pélvica e grau de espondilolistese. Foram analisados 310 pacientes que se dividem em 2 grupos: Anormal (210) e Normal (100). Com o objetivo de facilitar o entendimento, foram consideradas apenas 2 características: Raio pélvico e Grau de espondilolistese. Na Figura 1.3, a posição dos pacientes de acordo as variáveis e quadro clínico é apresentada. Além disso, dois círculos indicam a posição dos protótipos das classes Anormal e Normal, e um quadrado indica um paciente da classe Anormal.

Pela figura, o paciente visivelmente está tão próximo da primeira classe quanto da segunda. De fato, as distâncias do parciente para os protótipos das classes Anormal e Normal são, respectivamente, $558,84$ e $490,01$. Usando o tradicional método *Fuzzy C-Means*, os graus de pertinência seriam, respectivamente, $0,4972$ e $0,5328$, indicando que o paciente seria erroneamente classificado para a classe Normal, quando na verdade ele pertence à Anormal. Desta forma, teria-se um Falso Negativo pelo fato do diagnóstico ser negativo, embora a doença esteja presente. Por outro lado, analisando este conjunto através da disposição dos pacientes segundo cada variável separadamente obtem-se a seguinte figura:

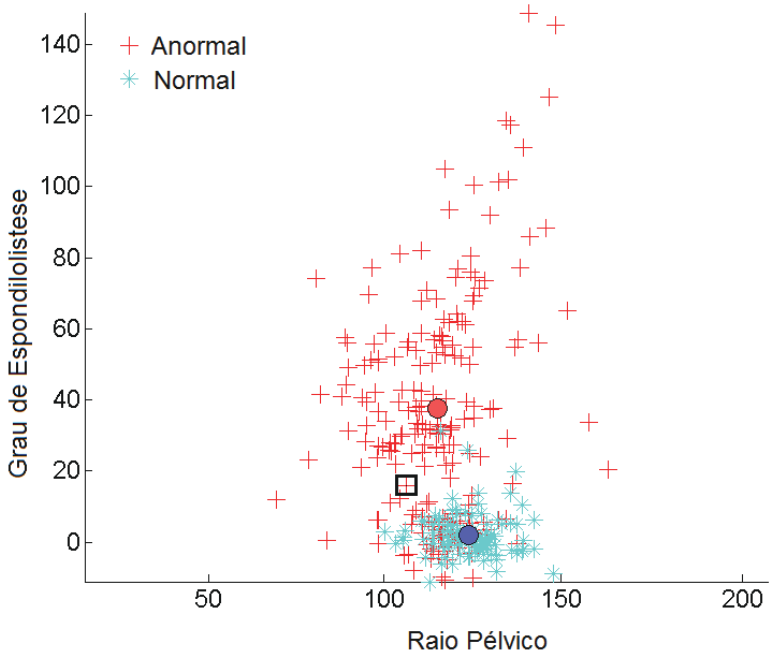

Figura 1.3: Conjunto de dados Coluna Vertebral de acordo com as variáveis "Raio pélvico" e "Grau de espondilolistese".

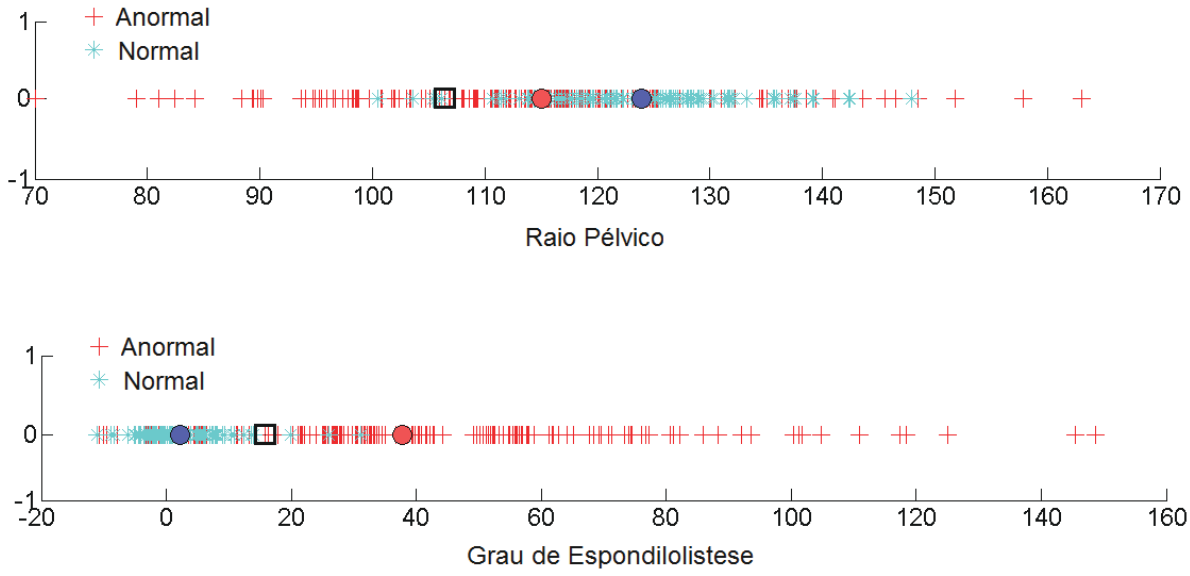

Figura 1.4: Análise de Coluna Vertebral com suas variáveis separadas: "Raio pélvico" em cima e "Grau de espondilolistese" embaixo.

Isto quer dizer que, segundo a variável "Raio pélvico", o paciente indicado pelo círculo à esquerda está mais próximo do protótipo da classe Anormal do que da classe Normal (imagem

mais acima). No que diz respeito à variável "Grau de espondilolistese" (imagem mais abaixo), o paciente está aproximadamente equidistante dos dois protótipos, ou seja está em uma zona de indecisão. Assim, a variável "Raio pélvico" é a que fornece uma informação mais relevante que a variável "Grau de espondilolistese" e indica fortemente que o paciente pertence à classe Anormal. Daí surge a necessidade de criar e utilizar graus de pertinência que também levem em conta a informação de cada variável separadamente. A partir deste tipo de grau de pertinência multivariado, duas vantagens podem ser apontadas: 1 - capacidade de interpretar a pertinência de cada objeto para um dado grupo segundo cada variável; 2 - obtenção de mais informação dos dados levando a uma maior qualidade de agrupamento.

1.2 Objetivos

Neste trabalho, serão apresentados métodos de agrupamento difuso onde os valores do grau de pertinência de cada objeto para um dado grupo é dado por um vetor p-dimensional, onde p é o número de variáveis. Isto é, será possível calcular quão similar um objeto é de um protótipo de acordo com cada variável. A principal proposta é proporcionar uma maior quantidade de informação para os algoritmos e obter uma melhor qualidade de agrupamento comparado com outros presentes na literatura. Os métodos podem ser divididos em duas categorias de acordo com a capacidade de encontrar partições com a presença ou não de dados aberrantes.

A primeira categoria é baseada no método *Fuzzy C-Means* (FCM) proposto por Bezdek (1981). Pelo fato do FCM ser sensível à presença de dados aberrantes, os métodos dessa categoria também degradam o desempenho em conjuntos com tal tipo de dados. Por outro lado, a segunda categoria é baseada no método *Possibilistic Fuzzy C-Means* (PFCM) proposto por Pal et al. (2005) e torna os métodos mais robustos a dados aberrantes. Além disso, também são propostos índices de interpretação para avaliar a qualidade do agrupamento para grupo e variável a partir da partição difusa obtida por cada método proposto. Nesse contexto, o intuito desse trabalho é responder duas perguntas (hipóteses) de pesquisa:

1. É possível melhorar a qualidade de agrupamento difuso através da utilização de graus de pertinência e de possibilidade cujos valores são calculados segundo cada variável?

2. É possível descrever a relevância de cada variável para cada grupo usando tais graus?

Com o objetivo de avaliar o desempenho dos métodos propostos e os presentes na literatura, um estudo comparativo destes métodos em relação ao agrupamento difuso usando o experimento Monte Carlo é realizado. Foram planejados experimentos com dados sintéticos e

reais e o índice de Rand difuso (FR) é usado para avaliar os métodos. Além disso, uma aplicação dos métodos propostos em dados biológicos também faz parte do objetivo deste trabalho.

1.3 Organização da Tese

A tese, segundo os capítulos, está organizada da seguinte forma:

Capítulo 2 - Métodos de Agrupamento Difuso

Neste capítulo, uma breve descrição do estado da arte da Análise de Agrupamento para a solução de problemas agrupamento difuso é apresentada.

Capítulo 3 - Métodos de Agrupamento Difuso Multivariado

Os métodos propostos serão apresentados neste capítulo da tese. Os métodos dizem respeito ao agrupamento de dados usando a abordagem difusa para encontrar uma partição. O primeiro método é baseado no algoritmo *Fuzzy C-Means* usando o grau de pertinência multivariado, o segundo e o terceiro são versões do primeiro usando dois tipos de ponderação. O quarto é a extensão do método *Possibilistic Fuzzy C-Means* onde tanto o grau de pertinência quanto o grau de possibilidade são multivariados. O quinto propõe a utilização de pesos para permitir que o método seja capaz de identificar grupos de diferentes formas e tamanhos. Finalmente, serão apresentadas ferramentas baseadas em medidas de dispersão para a interpretação da partição e grupos difusos obtidos por métodos multivariados. Estas ferramentas permitem medir a qualidade global, a homogeneidade dos grupos e o papel de diferentes variáveis no processo de formação do grupo.

Capítulo 4 - Experimentos e Resultados

Nesta parte da tese, serão tratados experimentos com os métodos propostos e os presentes na literatura para o problema de agrupamento difuso. Na primeira parte, os métodos serão avaliados através de diferentes configurações de dados artificiais usando experimento Monte Carlo. Na segunda parte, o desempenho será avaliado para conjuntos de dados reais usando o método de *Bootstrap*. Como métrica para a avaliação, será usado o índice de Rand difuso (FR) fazendo a comparação entre a partição *a priori* e a partição difusa obtida por cada método.

Capítulo 5 - Aplicação em Dados Biológicos

As motivações e os objetivos para a aplicação dos métodos propostos em dados biológicos serão relatados neste capítulo. A proposta é contribuir de forma prática para a comunidade científica através do uso de métodos de agrupamento difuso multivariado para extração e análise de informações em base de dados biológicos. Para isto, bancos de dados biológicos foram selecionados e usados para avaliar o desempenho dos métodos. São bibliotecas de informações científicas coletadas através de experimentos, publicados na literatura e armazenados usando técnicas computacionais. Além disso, uma estratégia para interpretar os graus de pertinência multivariados é introduzida nesse capítulo.

Capítulo 6 - Conclusões

Nesse capítulo, serão apresentadas as conclusões, trabalhos publicados referentes ao tema de pesquisa e propostas de trabalhos futuros.

2

MÉTODOS DE AGRUPAMENTO DIFUSO

Os que se encantam com a prática sem a ciência são como os timoneiros que entram no navio sem timão nem bússola, nunca tendo certeza do seu destino.

—LEONARDO DA VINCI

2.1 Introdução

O agrupamento de objetos similares para produzir uma categorização é uma das mais básicas abilidades dos seres-vivos (Everitt, 2011). O objetivo de ordenar elementos que compartilham propriedades em comum é importante para a identificação, por exemplo, de objetos comestíveis, venenozos, perigosos, entre outros. Assim como atividades básicas do ser humano, o agrupamento é também fundamental para muitos ramos da ciência. Em Biologia, por exemplo, existe a preocupação de dividir os diferentes animais ou plantas em grupos para melhor entendimento das funções biológicas. Em Arquilogia, pesquisadores usaram métodos de agrupamento para entender artefatos através de seus diferentes utilidades e assim poder dizer o quão pré-histórico um determinado povo viveu. Coompreender como expressões gênicas são seletivamente controladas é atualmente um grande desafio para a Bioinformática, desta forma técnicas de Análide de Agrupamente estão sendo empregadas para identificar grupos de genes com padrões de expressões similares e assim tentar responder qual grupo é responsável por uma determinada doença.

Em muitas situações, além de informar a qual grupo um determinado objeto pertence, é necessário entender quão similar este objeto está para todos os grupos da partição. Isto é, o objeto deve ter um grau de proximidade ou similaridades para os grupos (Cox, 2005). Isto deve-se ao fato que, em muitos problemas, existem informações com alguma forma de incerteza

ou imprecisão. Por exemplo, no estudo de identificação de regiões de cidades, a fronteira entre duas cidades nem sempre pode ser determinada de forma precisa e, assim, deve-se utilizar algum modelo que represente este comportamento (Petry e Robinson, 2005; Keller, 2005). Em 1967, Bill Wee escreveu a primeira tese de doutorado sobre reconhecimento de padrões usando a abordagem difusa (Keller, 2005). Ruspini (1969) definiu a matriz de pertinência e em 1970 descreveu o primeiro método de agrupamento difuso a partir de dados não rotulados. Em 1974, Dunn e Bezdek publicaram artigos sobre o modelo de agrupamento *Fuzzy C-Means*.

Neste capítulo, métodos para a solução do problema de agrupamento usando a abordagem difusa serão descritos. Inicialmente, o método de agrupamento mais amplamente conhecido *Fuzzy C-Means* (FCM) é apresentado na Seção 2.2 e serve como um conceito fundamental para os outros algoritmos. Além disso, versões do FCM presentes na literatura basedas em distâncias adaptativas também serão apresentadas nas Seções 2.3 e 2.4. A Seção 2.5 trata sobre uma adaptação do FCM para lidar com dados intervalares e ruidosos. Seção 2.6 descreve um método que usa tanto a abordagem difusa quanto a possibilística para ser mais robusto a *outliers*. Em seguida, será descrito uma aplicação do FCM para o agrupamento do tipo hierárquico na Seção 2.7. Finalmente, Seção 2.8 trata sobre o FCM com poderação por variável.

2.2 Fuzzy C-Means

O conceito de *fuzzy set* foi inicialmente explorado por Zadeh (1965) e posteriormente aplicado a métodos de agrupamento por Ruspini (1969). Recentemente, trabalhos com aplicações de *fuzzy set* em Análise de Agrupamento tem sido desenvolvidos em diferentes campos do conhecimento (Yang, 1993). Chen et al. elaborou diversos trabalhos a respeito da utilização do conceito de nebulosidade em variados métodos (Chen, 2009; Chen et al., 2010; Chen, 2010a; Yeh e Chen, 2010; Chen, 2010b).

A partir dos trabalhos de Zadeh, Dunn (1973) desenvolveu e Bezdek (1981) melhorou o método *Fuzzy C-Means* (FCM). Atualmente, o método FCM é o mais conhecido da abordagem difusa para particionamento de dados amplamente usado em Reconhecimento de Padrões. Sua principal característica é encontrar um conjunto de grupos tal que cada objeto possui graus de pertinência para cada classe. A partição final dependerá desses valores de graus de pertinência, desta forma, o cálculo desses valores torna-se importante para esta abordagem difusa. Uma vantagem deste método sobre o *K-Means*, por exemplo, é sua capacidade de encontrar resultados mais satisfatórios quando os dados são dispostos com sobreposição, isto é, objetos de diferentes classes possuem características similares. Entretanto, semelhante ao *K-Means*, o FCM pode levar a um mínimo local. O FCM busca minimizar a seguinte função objetivo:

$$J = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m d(\mathbf{x}_k, \mathbf{y}_i), \tag{2.1}$$

onde $d(\mathbf{x}_k, \mathbf{y}_i)$ é a distância Euclidiana entre o objeto $\mathbf{x}_k$ pertencente a um conjunto de objetos $\Omega = \{\mathbf{x}_1, \ldots, \mathbf{x}_k, \ldots, \mathbf{x}_n\}$ e o protótipo $\mathbf{y}_i$ do grupo C_i. Cada objeto é descrito por uma lista $\mathbf{x}_k = (x_k^1, \ldots, x_k^j, \ldots x_k^p)$, isto é, $\mathbf{x}_k \in \Re^p$ onde p é o numero de variáveis. Similarmente, $\mathbf{y}_i = (y_i^1, \ldots, y_i^j, \ldots y_i^p)$ onde cada protótipo pertence ao conjunto $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_i, \ldots, \mathbf{y}_c\}$. A distância é calculada como:

$$d(\mathbf{x}_k, \mathbf{y}_i) = \sum_{j=1}^{p} \left(x_k^j - y_i^j \right)^2. \tag{2.2}$$

O cálculo de J depende do protótipo $\mathbf{y}_i$, do grau de pertinência u_{ik} e do valor do parâmetro m (a influência deste para o cálculo do grau de pertinência é exemplificado na Figura 2.1). A cada iteração do algoritmo, os dois primeiros são recalculados buscando minimizar a função objetivo até que se atinja um mínimo local. A forma como o protótipo e o grau de pertinência são calculados é descrita a seguir.

$$\mathbf{y}_i = \frac{\sum_{k=1}^{n} (u_{ik})^m \mathbf{x}_k}{\sum_{k=1}^{n} (u_{ik})^m}. \tag{2.3}$$

Em Análise de Agrupamento, o grau de pertinência de um objeto k pertencer a um grupo C_i é expresso por $u_{ik} \in [0,1]$ sob a restrição $\sum_{i=1}^{c} u_{ik} = 1, k = 1, \ldots, n$ e seus valores são dispostos na matriz $\mathbf{U} = [u_{ik}]_{c \times n}$. O cálculo do grau de pertinência u_{ik}, de acordo com o tradicional *Fuzzy C-Means* de Zadeh, é feito da seguinte forma:

$$u_{ik} = \left[\sum_{h=1}^{c} \left\{ \frac{d(\mathbf{x}_k, \mathbf{y}_i)}{d(\mathbf{x}_k, \mathbf{y}_h)} \right\}^{\frac{1}{m-1}} \right]^{-1}. \tag{2.4}$$

Em resumo, o grau de pertinência u_{ik} de um objeto $\mathbf{x}_k$ para um grupo C_i segue as seguintes restrições:

1. $u_{ik} \in [0,1] \; \forall i, k$;
2. $0 < \sum_{k=1}^{n} u_{ik} < n \; \forall i$;
3. $\sum_{i=1}^{c} u_{ik} = 1 \; \forall k$.

O primeiro item garante que o grau de pertinência será um valor entre 0 e 1, qualquer que seja o objeto ou o grupo do agrupamento. A segunda restrição confirma a primeira onde o somatório será n se todos os n objetos tiverem uma chance de 100% de pertencerem a um

mesmo grupo, e 0 para os outros grupos. O último item afirma que a soma dos graus para todos os grupos fixando um objeto é necessariamente 1, qualquer que seja esse objeto.

O parâmetro m é um expoente porderador aplicado ao cálculo do grau de pertinência, esse valor influi no valor final de u_{ik} e determina o grau de fuzzificação: quando $m \to 1$ a função do grau de pertinência torna-se equivalente a Função Indicador e o método se torna do tipo *hard*; por outro lado, quando $m \to \infty$ a função é totalmente difusa. Devido à sua importância para o cálculo do grau de pertinência, diversos trabalhos foram propostos a respeito de m. Yu e Huang (2004) e Wu (2012), por exemplo, propuseram técnicas para tratar este parâmetro. A figura a seguir mostra o comportamento das curvas do grau de pertinência segundo a variação do somatório $\sum_{h=1}^{c} \frac{d(\mathbf{x}_k,\mathbf{y}_i)}{d(\mathbf{x}_k,\mathbf{y}_h)}$ para os valores de $m = 1,1$; $1,5$; 2; 3 e 5:

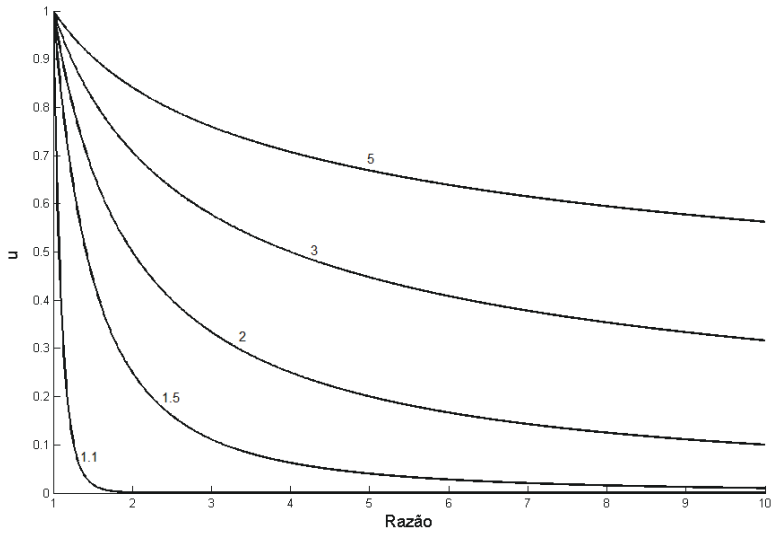

Figura 2.1: Variação do grau de pertinência em função do somatótio para diferentes valores de m.

O algoritmo do Método *Fuzzy C-Means* é descrito como a seguir:

Existem muitos artigos com trabalhos relacionados à teoria e à aplicação do algoritmo FCM como teoremas numéricos, teoremas estocásticos, processamento de imagens, estimação de parâmetros e muitos outros (Yang, 1993). Por exemplo, Hathaway et al. (2000) apresentaram versões do métodos FCM usando várias distâncias L_P, cujo principal objetivo é aumentar a robustez a *outliers*. Jajuga (2001) apresentou o algoritmo de agrupamento difuso baseado na norma L_1. Groenen e Jajuga (2001) apresentaram um novo agrupamento difuso com base na distância quadrada Minkowski. Recentemente, Lu et al. (2014) relatam uma versão melhorada do método FCM ao usar uma distância baseda no grau de densidade. Ferreira e De Carvalho (2014) propuseram um algoritmo basedo em kernel onde a distância entre objeto-protótipo é calculada

Algoritmo 1 FCM (Ω, c)

Requer: Conjunto de dados Ω e o número de grupos c;

1: Faça $\varepsilon > 0$. Fixe c, $2 \leq c \leq n$; fixe m, $1 < m < \infty$; fixe T; e estabeleça $\varepsilon > 0$; Inicialize aleatotiamente $u_{ik}(i = 1,\ldots,c$ e $k = 1,\ldots,n)$ do objeto k pertecente ao grupo C_i tal que $u_{ik} \in [0,1]$, $0 < \sum_{k=1}^{n} u_{ik} < n$ e $\sum_{i=1}^{c} u_{ik} = 1$, para todo $k \in \Omega$.;
2: $t \leftarrow 0$;
3: $J(t) \leftarrow 0$;
4: $J(t+1) \leftarrow \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m d(\mathbf{x}_k, \mathbf{y}_i)$;
5: **Enquanto** $|J(t) - J(t+1)| > \varepsilon$ **e** $t < T$ **Faça**
6: **Atualize matriz de protótipos Y**: Fixo o grau de pertinência u_{ik}, atualize os protótipos y_{ij} usando a Equação 2.3;
7: **Atualize matriz de graus de pertinênia U**: Fixo o protótipo y_{ij} $(i = 1,\ldots,c$ e $k = 1,\ldots,n)$, atualize o grau de pertinência u_{ik} usando a Equação 2.4;
8: $J(t) \leftarrow J(t+1)$;
9: $J(t+1) \leftarrow \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m d(\mathbf{x}_k, \mathbf{y}_i)$;
10: $t \leftarrow t+1$;
11: **Fim Enquanto**

Retorne Matrizes **Y** e **U**.

para cada variável com o uso de pesos que indicam a forma do grupo. Quanto às aplicações, Verma e Yadava (2014) usaram o método FCM para analisar polímeros e componentes orgânicos. Chatterjee et al. (2014) apresentaram uma abordagem para detecção de energia baseado no sensoriamento de espectro usando o FCM. Nas seções a seguir serão apresentados com mais detalhes outros métodos baseados na abordagem de *Fuzzy Sets* e agrupamento difuso.

2.3 Fuzzy C-Means com Ponderação por Variável

O método proposto por Keller e Klawonn (2000) possui uma função objetivo baseada em técnica de agrupamento difuso que leva em conta um peso que mede a influência de cada variável do conjunto de dados para cada grupo da partição (aqui esse método será abreviado como FCM-C[1]). Portanto, o método não é somente adequado para encontratar agrupamento difuso, mas também dá mais informação a respeito da cada variável na detecção de grupos. De acordo com os autores, o método pode ser considerado uma generalização do FCM. A função do *Fuzzy C-Means* com Ponderação por Variável é dado como a seguir:

$$J = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m d_{ik}, \tag{2.5}$$

onde c é o número de grupos, n o número de objetos, m é o parâmetro de funzzificação e d_{ik} é a distância Euclidiana dada como:

$$d_{ik} = \sum_{j=1}^{p} (\alpha_{ij})^t (x_{jk} - y_{ij})^2, \tag{2.6}$$

onde x_{jk} e y_{ij} são as coordenadas dos vetores $\mathbf{x}_k$ e $\mathbf{y}_i$, respectivamente. O número de variáveis é dado por p. Além disso, α_{ij} é o peso que determina a influência da variável j para o grupo i. Este parâmetro pode ser fixo ou adaptativo de acordo com a execução do algoritmo e segue a seguinte restrição:

$$\sum_{j=1}^{p} \alpha_{ij} = a, \forall i \in \{1, \ldots, c\}. \tag{2.7}$$

de forma que $a \in \Re$ é um valor constante, por exemplo, $a = 1$ ou $a = c$. O valor de $t \in \Re_{\geq 1}$ é uma influência sobre o peso α_{ij} similarmente como o valor de m tem sobre o grau de pertinência u_{ik}. Quando $t \to 1$, o peso α_{ij} tende a ser 0 (não tem influência) ou 1 (não existe distinção entre as variáveis). Por outro lado, quando $t \to \infty$ todos as variáveis recebem a mesma influência no agrupamento, isto é, $\alpha_{ij} = \frac{1}{a}$ para todo i e j. O valor de α_{ij} que minimiza a função objetivo 2.5 é calculado como na equação a seguir:

$$\alpha_{ij} = \left[\sum_{r=1}^{p} \left(\frac{\sum_{k=1}^{n} (u_{ik})^m (x_{jk} - y_{ij})^2}{\sum_{k=1}^{n} (u_{ik})^m (x_{rk} - y_{ir})^2} \right)^{\frac{1}{t-1}} \right]^{-1} \tag{2.8}$$

O protótipo e o grau de pertinência que também minimizam a função objetivo é dado como nas equações 2.3 e 2.4, respectivamente.

O algoritmo do Método FCM-C[1] é descrito como a seguir:

2.4 Fuzzy C-Means com Distância Adaptativa

Semelhantemente ao método *Fuzzy C-Means* com Ponderação por Variável como apresentado na seção anterior, uma versão do tradicional FCM para dados clássicos foi proposta por De Carvalho et al. (2006a) onde os métodos usam distância adaptativa proposta por Diday (Diday e Simon, 1976; Diday e Gouvaert, 1976). A principal ideia da distância adaptativa é usar ponderadores que retratem as dispersões dos grupos e que mudem a cada iteração do algoritmo. Assim, os métodos tornam-se capazes de diferenciar grupos de diferentes tamanhos e formas, aumentando a qualidade da partição quando comparado com o tradicional FCM. Os pesos que ponderam as distâncias podem ter duas abordagens: global ou por classe. Para cada abordagem, existe uma matriz de covariância que pode ser de dois tipos: cheia ou diagonal. Os métodos apresentados nessa seção são para o caso em que a matriz de covariância é diagonal.

Algoritmo 2 FCM-C[1] (Ω, c)

Requer: Conjunto de dados Ω e o número de grupos c;

1: Faça $\varepsilon > 0$. Fixe c, $2 \leq c \leq n$; fixe m, $1 < m < \infty$; fixe T; e estabeleça $\varepsilon > 0$; Inicialize aleatotiamente $u_{ik}(i = 1,\ldots,c \text{ e } k = 1,\ldots,n)$ do objeto k pertecente ao grupo C_i tal que $u_{ik} \in [0,1]$, $0 < \sum_{k=1}^{n} u_{ik} < n$ e $\sum_{i=1}^{c} u_{ik} = 1$, para todo $k \in \Omega$. Faça $\alpha_{ij} = 1$, para todo i e j;
2: $t \leftarrow 0$;
3: $J(t) \leftarrow 0$;
4: $J(t+1) \leftarrow \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m d_{ik}$;
5: **Enquanto** $|J(t) - J(t+1)| > \varepsilon$ **e** $t < T$ **Faça**
6: **Atualize os protótipos**: Fixo o grau de pertinência u_{ik} e o peso α_{ij}, atualize os protótipos y_{ij} usando a Equação 2.3;
7: **Atualize os graus de pertinênia**: Fixo o protótipo y_{ij} $(i = 1,\ldots,c \text{ e } k = 1,\ldots,n)$ e o peso α_{ij}, atualize o grau de pertinência u_{ik} usando a Equação 2.4;
8: **Atualize os pesos**: Fixo o protótipo y_{ij} $(i = 1,\ldots,c \text{ e } k = 1,\ldots,n)$ e o grau de pertinência u_{ik}, atualize o peso α_{ij} usando a Equação 2.8;
9: $J(t) \leftarrow J(t+1)$;
10: $J(t+1) \leftarrow \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m d_{ik}$;
11: $t \leftarrow t+1$;
12: **Fim Enquanto**

Retorne Partição difusa.

A distância global com matriz diagonal usa pesos que variam de acordo com a dispersão de cada variável sob a suposição de que todos os grupos sejam indiferentes quanto a essa dispersão, isto é, não há diferenciação de pesos para cada grupo. O critério do método *Fuzzy C-Means* com Distância Adaptativa Global (aqui abreviado FCM-S) é dado como a seguir:

$$J = \sum_{j=1}^{p} \lambda_j \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m (x_{jk} - y_{ij})^2. \tag{2.9}$$

onde o protótipo de que minimiza esta função objetivo é calculado seguindo a seguinte equação:

$$\mathbf{y}_i = \frac{\sum_{k=1}^{n}(u_{ik})^m \mathbf{x}_k}{\sum_{k=1}^{n}(u_{ik})^m}. \tag{2.10}$$

O grau de pertinência que minimiza o critério é similarmente calculado como no tradicional FCM. A principal diferença é que leva em conta o peso λ_j:

$$u_{ik} = \left[\sum_{h=1}^{c}\left\{\frac{\sum_{j=1}^{p}\lambda_j(x_{jk} - y_{ij})^2}{\sum_{j=1}^{p}\lambda_j(x_{jk} - y_{hj})^2}\right\}^{\frac{1}{m-1}}\right]^{-1}, \tag{2.11}$$

O peso λ_j, por sua vez que minimiza a função objetivo segundo a restrição $\prod_{j=1}^{p}\lambda_j = 1$ é dado como a seguinte expressão:

$$\lambda_j = \frac{\left\{\prod_{h=1}^{p}\left[\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m(x_{hk}-y_{ih})^2\right]\right\}^{\frac{1}{p}}}{\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m(x_{jk}-y_{ij})^2} \tag{2.12}$$

A distância por classe também varia a cada iteração do algoritmo, diferentemente do tradicional FCM, onde a distância é fixa. Além disso, existe uma ponderação por variável e grupo, de forma que o método que usa esse tipo de distância pode identificar conjunto de elementos de diferentes formas e tamanhos. Assim como a distância adaptativa global, a por classe pode ser diferenciada de acordo com o tipo de matriz que usa: cheia ou diagonal. Para o método *Fuzzy C-Means* com Distância Adaptativa por Classe e matriz diagonal (aqui abreviado FCM-C^2), o problema é encontrar um conjunto de protótipos, uma matriz difusa e vetores de pesos tais que minimizem iterativamente a função objetivo descrita como:

$$J = \sum_{j=1}^{p}\sum_{i=1}^{c}\lambda_i^j\sum_{k=1}^{n}(u_{ik})^m(x_{jk}-y_{ij})^2. \tag{2.13}$$

onde o protótipo que minimiza esta função objetivo é igualmente calculado como descrito na Equação 2.10. O grau de pertinência também é similarmente calculado como apresentado no método anterior, entretanto leva em conta o peso por variável e grupo λ_i^j:

$$u_{ik} = \left[\sum_{h=1}^{c}\left\{\frac{\sum_{j=1}^{p}\lambda_i^j(x_{jk}-y_{ij})^2}{\sum_{j=1}^{p}\lambda_h^j(x_{jk}-y_{hj})^2}\right\}^{\frac{1}{m-1}}\right]^{-1}. \tag{2.14}$$

O peso por variável e grupo λ_i^j que minimiza o critério do método sob a restrição $\prod_{j=1}^{p}\lambda_i^j = 1$ para todo $i = 1,\ldots,c$ é calculado como a seguir:

$$\lambda_i^j = \frac{\left\{\prod_{h=1}^{p}\left[\sum_{k=1}^{n}(u_{ik})^m(x_{hk}-y_{ih})^2\right]\right\}^{\frac{1}{p}}}{\sum_{k=1}^{n}(u_{ik})^m(x_{jk}-y_{ij})^2} \tag{2.15}$$

O algoritmo comum aos métodos FCM-S e FCM-C^2 é descrito como a seguir:

2.5 Fuzzy C-Means Robusto

Este tipo de método tem duas principais características proposto por D'Urso e Giordani (2006). O primeiro ponto importante é a extensão do tradicional FCM para dados simbólicos do tipo intervalo. Uma variável simbólica pode representar o resumo de banco de dados a partir de um valor categórico, ou a síntese de informações que possuem características em comum e essa representação pode ser feita utilizando listas, intervalos, distribuições ou funções mostrando a

Algoritmo 3 FCM-S/FCM-C^2 (Ω, c)

Requer: Conjunto de dados Ω e o número de grupos c;

1: Faça $\varepsilon > 0$. Fixe c, $2 \leq c \leq n$; fixe m, $1 < m < \infty$; fixe T; e estabeleça $\varepsilon > 0$; Inicialize aleatotiamente $u_{ik} (i = 1, \ldots, c$ e $k = 1, \ldots, n)$ do objeto k pertecente ao grupo C_i tal que $u_{ik} \in [0,1]$, $0 < \sum_{k=1}^{n} u_{ik} < n$ e $\sum_{i=1}^{c} u_{ik} = 1$, para todo $k \in \Omega$. Faça $\lambda_j = 1$ ou $\lambda_i^j = 1$ $(i = 1 \ldots, c$ e $j = 1, \ldots, p)$;
2: $t \leftarrow 0$;
3: $J(t) \leftarrow 0$;
4: $J(t+1) \leftarrow$ critério 2.9 para FCM-S ou 2.13 para FCM-C^2;
5: **Enquanto** $|J(t) - J(t+1)| > \varepsilon$ **e** $t < T$ **Faça**
6: **Atualize matriz de protótipos Y**: Fixo o grau de pertinência u_{ik} e o peso λ_j (FCM-S) ou λ_i^j (FCM-C^2), atualize os protótipos y_{ij} usando a Equação 2.10;
7: **Atualize os pesos**: Fixo o grau de pertinência u_{ik} e o protótipo y_{ij}, atualize os pesos λ_j usando a Equação 2.12 (FCM-S) ou λ_i^j usando a Equação 2.15 (FCM-C^2);
8: **Atualize matriz de graus de pertinênia U**: Fixo o protótipo y_{ij} $(i = 1, \ldots, c$ e $k = 1, \ldots, n)$ e o peso λ_j (FCM-S) ou λ_i^j (FCM-C^2), atualize o grau de pertinência u_{ik} usando a Equação 2.11 (FCM-S) ou Equação 2.14 (FCM-C^2);
9: $J(t) \leftarrow J(t+1)$;
10: $J(t+1) \leftarrow$ critério 2.9 para FCM-S ou 2.13 para FCM-C^2;
11: $t \leftarrow t+1$;
12: **Fim Enquanto**

Retorne Matrizes **Y**, **U** e vetores de pesos.

variabilidade contida na informação original. O objetivo da Análise Simbólica de Dados (ADS) é generalizar mineração de dados e estatística para unidades de mais alto nível descritas por dados simbólicos (Diday e Noirhomme-Fraiture, 2008). A principal motivação de ADS é reduzir a dimensão da tabela de dados original e criar objetos simbólicos a partir destes dados de forma que a perda de informação seja mínima, haja uma diminuição do espaço de armazenamento e um aumento da velocidade de processamento deste dados simbólicos. Outra característica deste método é a sua robustez a dados ruidosos, mostrando ser mais eficiente em conjuntos contendo tais tipos de dados.

No método FCM robusto para dados intervalares, o algoritmo busca minimizar a seguinte função objetivo:

$$J = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m d^2(\mathbf{x}_k, \mathbf{y}_i) + \sum_{k=1}^{n} \delta^2 \left(1 - \sum_{i=1}^{c} u_{ik}\right)^m, \tag{2.16}$$

onde $d^2(\mathbf{x}_k, \mathbf{y}_i)$ mede a dissimilaridade entre o objeto $\mathbf{x}_k$ e o protótipo $\mathbf{y}_i$ do grupo C_i e é dada por:

$$d^2(\mathbf{x}_k, \mathbf{y}_i) = \left[\|\mathbf{c}_i - \mathbf{c}_k\|^2 + \|\mathbf{r}_i - \mathbf{r}_k\|^2\right] = \sum_{j=1}^{p} \left[(c_i^j - c_k^j)^2 + (r_i^j - r_j^j)^2\right]. \tag{2.17}$$

A dissimilaridade leva em consideração a distância entre os centros e as amplitudes dos objetos e protótipos para todas as variáveis intervalares. Um intervalo representa a variabilidade de um conjunto de dados quantitativos onde os limites representam o mínimo e o máximo dos valores desta observação. A partir do intervalo, duas informações podem ser obtidas: o centro e a amplitude. A relação entre os limites de um intervalo para a variável k de um objeto $\mathbf{x}_j$ é dado a seguir:

$$[a_k^j, b_k^j] \rightarrow a_k^j = c_k^j - r_k^j \text{ e } b_k^j = c_k^j + r_k^j, \tag{2.18}$$

onde c_k^j é o centro e r_k^j é o range. A partir desta relação, estes valores podem ser obtidos da seguinte forma:

$$c_k^j = \frac{a_k^j + b_k^j}{2} \text{ e } r_k^j = \frac{b_k^j - a_k^j}{2}. \tag{2.19}$$

Assim, cada objeto $\mathbf{x}_k$ pode ser representado por um par de vetores, isto é, $\mathbf{x}_k = (\mathbf{c}_k, \mathbf{r}_k)$, onde $\mathbf{c}_k = (c_k^1, c_k^2, \ldots, c_k^p)$ e $\mathbf{r}_k = (r_k^1, r_k^2, \ldots, r_k^p)$.

Outra importante característica deste método é a restrição do grau de pertinência. A proposta é que o algoritmo calcule estes valores de forma que $\sum_{i=1}^{c} u_{ij} \leq 1$. Assim, existe um tratamento diferenciado para conjuntos com dados ruidosos e valores atípicos uma vez que elementos deste tipo receberão um valor de grau de pertinência menor que outros padrões da base de dados. Para isto, o grau de pertinência é calculado como a seguir:

$$u_{ik} = \left\{ \sum_{h=1}^{c} \left[\frac{d^2(\mathbf{x}_k, \mathbf{y}_i)}{d^2(\mathbf{x}_k, \mathbf{y}_h)} \right]^{\frac{1}{m-1}} + \left[\frac{d^2(\mathbf{x}_k, \mathbf{y}_i)}{\delta^2} \right]^{\frac{1}{m-1}} \right\}^{-1}, \tag{2.20}$$

onde cada protótipo $\mathbf{y}_i$, assim como objeto, pode ser representado por um par de vetores $(\mathbf{c}_i, \mathbf{r}_i)$. O vetor de centros $\mathbf{c}_i = (c_i^1, c_i^2, \ldots, c_i^p)$ é dado por:

$$\mathbf{c}_i = \frac{\sum_{k=1}^{n} (u_{ik})^m \mathbf{c}_k}{\sum_{k=1}^{n} (u_{ik})^m} \tag{2.21}$$

e o vetor de amplitudes $\mathbf{r}_i = (r_i^1, r_i^2, \ldots, r_i^p)$ é calculado como:

$$\mathbf{r}_i = \frac{\sum_{k=1}^{n} (u_{ik})^m \mathbf{r}_k}{\sum_{k=1}^{n} (u_{ik})^m}. \tag{2.22}$$

Finalmente, a constante δ garante que o grau de pertinência seja menor do que 1 e pode ser interpretada como uma "distância de ruido". A constante é dada por:

$$\delta = \sqrt{\frac{\sum_{k=1}^{n}\sum_{j=1}^{p} c_k^j}{np}} + \sqrt{\frac{\sum_{k=1}^{n}\sum_{j=1}^{p} r_k^j}{np}} \tag{2.23}$$

2.6 Possibilistic Fuzzy C-Means

Em muitas aplicações de dados reais, existem elementos que podem ser considerados ruidosos por apresentarem um comportamento diferenciado em relação ao restante dos dados. Existem muitos motivos que proporcionam o seu surgimento, por exemplo, falha no armazenamento dos dados por máquinas ou humanos e erro no processamento da informação (como em reconhecimento de voz). Além disso, grupos de pessoas, animais ou plantas podem possuir indivíduos que se comportam diferentemente do normal tornam-se valores atípicos (*outliers*) e em geral apresentam um afastamento em relação aos demais do grupo. Tanto dados ruidosos quanto *outliers* podem prejudicar o desempenho dos métodos de Reconhecimento de Padrões, inclusive o tradicional FCM.

O método *Possibilistic C-Means* (PCM) proposto por Krishnapuram e Keller (1993) é o mais popular para tentar resolver o problema de agrupamento não-rígido de forma que seja mais robusto a dados aberrantes. Deste método surge o conceito de grau de possibilidade onde a ideia é similar ao grau de pertinência, entretanto não existe a restrição onde o somatório de seus valores para todas as classes é necessariamente igual a 1. Desta forma, os algoritmos que usam a abordagem possibilística são mais flexíveis. Entretanto, PCM apresenta problemas como encontrar grupos coincidentes (os protótipos são similares) e não encontra resultados tão bons quanto os que são econtrados quando se utiliza o FCM. Assim, Pal et al. (1997) propuseram uma combinação entre o FCM o PCM (chamado de *Fuzzy Possibilistic C-Means* ou FPCM) justificando a necessidade de usar ambos os graus de possibilidade e pertinência, mas o método assemelha-se ao FCM com grandes conjuntos de dados. Finalmente, Pal et al. (2005) apresentaram o *Possibilistic Fuzzy C-Means* (PFCM) com o intuito de suprir as deficiências de FCM, PCM e FPCM.

O objetivo do algoritmo é minimizar a função objetivo a qual é calculada como descrita na equação a seguir:

$$J = \sum_{i=1}^{c}\sum_{k=1}^{n} \left\{ [a(u_{ik})^m + b(t_{ik})^\eta] d_{ik} + \pi_i (1 - t_{ik})^\eta \right\}, \tag{2.24}$$

onde d_{ik} é a distância Euclidiana entre o objeto k e o protótipo do grupo i. Os valores $a, b > 0$ são definidos *a priori* e indicam a importância do grau de pertinência e possibilidade, respectivamente, para o agrupamento. O parâmetro η diz respeito à importância de considerar os graus

de possibilidade t_{ik} para o agrupamento (normalmente é igual a 2). O valor de π_i serve para balancear os dois termos da equação, garantir que o critério é calculado buscando maximizar os valores de t_{ik} e pode ser escolhido pelo usuário. Entretanto, os autores sugerem usar a média ponderada das distâncias dada pela seguinte equação:

$$\pi_i = K \frac{\sum_{k=1}^{n} (u_{ik})^m d_{ik}}{\sum_{k=1}^{n} (u_{ik})^m}. \tag{2.25}$$

O valor de K é normalmente igual a 1 (Pal et al., 2005). O protótipo que minimiza a função objetivo leva em conta tanto o grau de pertinência quanto o de possibilidade e é calculado pela equação a seguir:

$$y_{ij} = \frac{\sum_{k=1}^{n} [a(u_{ik})^m + b(t_{ik})^\eta] x_{ik}}{\sum_{k=1}^{n} [a(u_{ik})^m + b(t_{ik})^\eta]}. \tag{2.26}$$

O grau de pertinência é similamente calculado como no método FCM sob a restrição $\sum_{i=1}^{c} u_{ik} = 1$. É dado pela seguinte expressão:

$$u_{ik} = \left[\sum_{h=1}^{c} \left(\frac{d_{ik}}{d_{hk}} \right)^{\frac{2}{m-1}} \right]^{-1}. \tag{2.27}$$

Finalmente, o grau de possibilidade diferencia-se do grau de pertinência porque não existe a restrição do somatório. Além disso, tem influência do valor do parâmetro b: quanto maior o valor de b, mais rapidamente o valor de t_{ik} decairá, isto é, supõe-se que o conjunto de dados possui objetos aberrantes e, portanto, o grau de possibilidade deve decrescer mais rapidamente em função da distância. Este grau é calculado como descrito na equação a seguir:

$$t_{ik} = \left[1 + \left(b \frac{d_{ik}}{\pi_i} \right)^{\frac{2}{(\eta-1)}} \right]^{-1}. \tag{2.28}$$

O algoritmo do método PFCM é descrito como a seguir:

2.7 Hierarchical Fuzzy Clustering

Nos algoritmos de agrupamento hierárquicos, os dados não são particionados em um determinado número de grupos em um único passo. Nestes métodos, o objetivo é obter uma informação mais completa sobre o conjunto de objetos por meio de um conjunto hierarquicamente aninhado de partições. Em taxonomia, por exemplo, um objeto pode pertencer sucessivamente a uma espécie, a um gênero, uma família, uma ordem, etc. O agrupamento consiste de uma

Algoritmo 4 PFCM (Ω, c)

Requer: Conjunto de dados Ω e o número de grupos c;

1: Faça $\varepsilon > 0$. Fixe c, $2 \leq c \leq n$; fixe m, $1 < m < \infty$; fixe T; e estabeleça $\varepsilon > 0$; Inicialize aleatotiamente $u_{ik}(i = 1, \ldots, c$ e $k = 1, \ldots, n)$ do objeto k pertecente ao grupo C_i tal que $u_{ik} \in [0,1]$, $0 < \sum_{k=1}^{n} u_{ik} < n$ e $\sum_{i=1}^{c} u_{ik} = 1$, para todo $k \in \Omega$. Inicialize aleatoriamente t_{ik} do objeto k pertecente ao grupo C_i. Calcule o valor de π_i $(i = 1 \ldots, c)$ usando a Equação 2.25;
2: $t \leftarrow 0$;
3: $J(t) \leftarrow 0$;
4: $J(t+1) \leftarrow \sum_{k=1}^{n} \{[a(u_{ik})^m + b(t_{ik})^\eta]\, d_{ik} + \pi_i (1 - t_{ik})^\eta\}$;
5: **Enquanto** $|J(t) - J(t+1)| > \varepsilon$ **e** $t < T$ **Faça**
6: **Atualize os protótipos**: Fixo o grau de pertinência u_{ik} e o grau de possibilidade t_{ik}, atualize os protótipos y_{ij} usando a Equação 2.26;
7: **Atualize os graus de pertinência**: Fixo o protótipo y_{ij} $(i = 1, \ldots, c$ e $k = 1, \ldots, n)$ e o grau de possibilidade t_{ik}, atualize o grau de pertinência u_{ik} usando a Equação 2.27;
8: **Atualize os graus de possibilidade**: Fixo o grau de pertinência u_{ik} e o protótipo y_{ij}, atualize os graus de possibilidade t_{ik} usando a Equação 2.28;
9: $J(t) \leftarrow J(t+1)$;
10: $J(t+1) \leftarrow \sum_{k=1}^{n} \{[a(u_{ik})^m + b(t_{ik})^\eta]\, d_{ik} + \pi_i (1 - t_{ik})^\eta\}$;
11: $t \leftarrow t + 1$;
12: **Fim Enquanto**

Retorne Partição difusa.

série de partições, as quais podem possuir apenas uma classe contendo todos os indivíduos, ou n grupos, cada um contendo apenas um elemento.

As técnicas de agrupamento hierárquicos podem ser divididos em duas categorias: algoritmos aglomerativos e algoritmos divisivos. Os aglometativos reduzem os dados em um único grupo contendo todos os elementos, enquanto que a técnica divisiva irá dividir o conjunto de entrada em n grupos, onde cada um contém apenas um indivíduo. Com isso, surge a necessidade de decidir em qual estágio do algoritmo irá parar, uma vez que é preciso encontrar uma solução com um número "ótimo" de classes.

Um método de agrupamento hierárquico aglometativo produz um série de partições do conjunto de dados, $C_n, C_{n-1}, \ldots, C_1$. O primeiro, C_n, consiste de n classes com um único membro. O último, C_1, consiste de um único grupo contendo todos os n elementos. Para executar o método hierárquico, comumente é utilizado uma matriz de dissimilaridade M entre pares de grupos. No início da execução do algoritmo aglomerativo, como todos os grupos são constituídos de 1 elemento, a matriz de dissimilaridade entre os grupos é a mesma obtida quando calcula-se essa métrica por pares de elementos.

Os métodos hierárquicos divisivos são caracterizados por reunir todos os n elementos da base em uma única classe e progressivamente separá-los de modo que no final da execução tenham-se n classes com apenas um elemento em cada uma. Os algoritmos divisivos podem

ser divididos em dois principais grupos: monotéticos (*monothetics*), que dividem o conjunto de dados levando-se em conta apenas um único atributo, e politéticos (*polythetics*), cujas divisões são realizadas a partir dos diversos atributos que um padrão pode ter.

No método *Hierarchical Fuzzy Clustering* (aqui abreviado por HFC) proposto por Torra (2005), existe uma combinação entre algoritmo particional difuso com o método hierárquico do tipo aglomerativo. Inicialmente, o algoritmo começa criando uma partição difusa dos dados aplicando o FCM. Ao final desse processo, uma matriz de pertinência e um conjunto de protótipos são definidos. Então esse processo é aplicado iterativamente para construir o agrupamento hierárquico de forma que o resultado da partição difusa no passo anterior será considerado como um novo conjunto de grupos que será agrupado usando o FCM.

No início, o número de grupos escolhido c deve ser grande o suficiente para garantir que o método hierárquico difuso terá muitas folhas no começo da execução. O FCM é executado similarmente como o algoritmo descrito na seção anterior. A principal diferença consiste na forma como a distância é calculada. Enquanto no FCM a distância é calculada entre objeto e protótipo, o algoritmo hierárquico difuso mede a dissimilaridade entre dois conjuntos difusos. O cálculo do grau de pertinência entre dois conjuntos difusos r e s é dado como a seguir:

$$u_{rs} = \left[\sum_{h=1}^{c'} \left\{ \frac{d(\mathbf{v}_r, \mathbf{v}_s)}{d(\mathbf{v}_r, \mathbf{v}_h)} \right\}^{\frac{1}{m-1}} \right]^{-1}, \tag{2.29}$$

onde $\mathbf{v}_r$ é o protótipo do grupo sendo agrupado e $\mathbf{v}_h$ são os protótipos dos grupos que estão sendo construidos pelo FCM na iteração atual. Similarmente c' é o número de grupos que o FCM está usando na iteração atual para encontrar os grupos. Outro componente importante a ser calculado é o novo protótipo:

$$\mathbf{v}_s = \frac{\sum_{k=1}^{n} n(u_{sk})^m \mathbf{x}_k}{\sum_{k=1}^{n} (u_{sk})^m} \tag{2.30}$$

O algoritmo do Método HFC é descrito como a seguir:

Segundo experimentos com o conjunto de dados Iris que os autores fizeram, o valor do parâmetro de fuzzificação m deve ser pequeno depois da primeira iteração do algoritmo, uma vez que o nível de nebulosidade entre os grupos aumenta devido à forma como os protótipos são calculados. Outra característica particular deste método é que existem diferentes graus de pertinência dependendo da altura da árvore e os graus dos grupos de um dado nível são inferiores que os graus dos seus sub-grupos.

Algoritmo 5 HFC(Ω)

Requer: Conjunto de dados Ω;
1: Defina $c \approx n$;
2: $P \leftarrow FCM(\Omega, c)$;
3: Defina $c' < c$;
4: **Enquanto** $c' \neq 1$ **Faça**
5: $P' \leftarrow FCM(P, c)$;
6: Faça a ligação entre P e P';
7: Faça $P = P'$;
8: Atualize o valor de c';
9: **Fim Enquanto**
Retorne P.

2.8 Fuzzy C-Means Ponderado

Mais recentemente, Nazari et al. (2013) propuseram uma melhoria do FCM utilizando pesos para ponderar automaticamente variáveis de acordo com sua importância no agrupamento. O Método *Fuzzy C-Means* Ponderado (abreviado por W-FCM) é uma extensão do método *WK-Means* proposto por Huang et al. (2005) que trata da ponderação no *K-Means*. O objetivo do método W-FCM é minimizar a seguinte função objetivo:

$$J = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \left(d_{ik}^{(w)}\right)^2, \tag{2.31}$$

onde a dissimilaridade entre o objeto x_{kj} e o protótipo y_{ij} é calculado como apresentada na equação a seguir:

$$d_{ik}^{(w)} = \sqrt{\sum_{j=1}^{p} w_j^{\beta} (x_{kj} - y_{ij})^2}. \tag{2.32}$$

Essa distância é similar à distância Euclidiana, exceto pelo fato dela considerar um peso w_j para cada variável. O protótipo, por sua vez, que minimiza a função objetivo, é dado como a seguir:

$$\mathbf{y}_i = \frac{\sum_{k=1}^{n} (u_{ik})^m \mathbf{x}_k}{\sum_{k=1}^{n} (u_{ik})^m}. \tag{2.33}$$

O grau de pertinência no método W-FCM também é calculado similarmente ao FCM, mas a principal diferença é que o W-FCM considera que as distâncias entre objetos e protótipos podem variar de acordo com cada variável. O grau de pertinência no W-FCM que minimiza a função objetivo é dado pela seguinte equação:

$$u_{ik} = \left[\sum_{h=1}^{c} \left\{ \frac{\left(d_{ik}^{(w)}\right)^2}{\left(d_{hk}^{(w)}\right)^2} \right\}^{\frac{1}{m-1}} \right]^{-1} . \tag{2.34}$$

O peso que tem o intuito de quantificar a importância de cada variável é computada de acordo com a função a seguir. Diferentemente de outros métodos (como o FCM-C[1] e FCM-C[2], por exemplo), o peso no W-FCM considera que não existe diferença entre grupos, isto é, qualquer que seja o grupo, o peso será o mesmo para uma dada variável.

$$w_j = \left[\sum_{t=1}^{p} \left\{ \frac{D_j}{D_t} \right\} \right]^{\frac{1}{\beta-1}} . \tag{2.35}$$

Uma vez que o peso não varia com os grupos, ele considera uma distância global calculada da seguinte forma:

$$D_j = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}(x_{kj} - y_{ij})^2 \tag{2.36}$$

O algoritmo do Método W-FCM é descrito como a seguir:

Algoritmo 6 W-FCM-C (Ω, c)

Requer: Conjunto de dados Ω e o número de grupos c;

1: Faça $\varepsilon > 0$. Fixe c, $2 \leq c \leq n$; fixe m, $1 < m < \infty$; fixe T; e estabeleça $\varepsilon > 0$; Inicialize aleatotiamente $u_{ik}(i = 1,\ldots,c$ e $k = 1,\ldots,n)$ do objeto k pertecente ao grupo C_i tal que $u_{ik} \in [0,1]$, $0 < \sum_{k=1}^{n} u_{ik} < n$ e $\sum_{i=1}^{c} u_{ik} = 1$, para todo $k \in \Omega$. Faça $w_j = 1$, para todo j;
2: $t \leftarrow 0$;
3: $J(t) \leftarrow 0$;
4: $J(t+1) \leftarrow \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \left(d_{ik}^{(w)}\right)^2$;
5: **Enquanto** $|J(t) - J(t+1)| > \varepsilon$ **e** $t < T$ **Faça**
6: **Atualize os protótipos**: Fixo o grau de pertinência u_{ik} e o peso w_j, atualize os protótipos y_{ij} usando a Equação 2.33;
7: **Atualize os graus de pertinênia**: Fixo o protótipo y_{ij} $(i = 1,\ldots,c$ e $k = 1,\ldots,n)$ e o peso w_j, atualize o grau de pertinência u_{ik} usando a Equação 2.34;
8: **Atualize os pesos**: Fixo o protótipo y_{ij} $(i = 1,\ldots,c$ e $k = 1,\ldots,n)$ e o grau de pertinência u_{ik}, atualize o peso w_j usando a equações 2.35 e 2.36;
9: $J(t) \leftarrow J(t+1)$;
10: $J(t+1) \leftarrow \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \left(d_{ik}^{(w)}\right)^2$;
11: $t \leftarrow t+1$;
12: **Fim Enquanto**

Retorne Partição difusa.

3

MÉTODOS DE AGRUPAMENTO DIFUSO MULTIVARIADO

A mente que se abre a uma nova idéia jamais voltará ao seu tamanho original.

—ALBERT EINSTEIN

3.1 Introdução

Para os métodos descritos, a seguir considere as seguintes definições.

Seja $\Omega = \{1,\ldots,k,\ldots,n\}$ um conjunto de n objetos listados por k. Cada objeto k é representado por um vetor de variáveis quantitativas $\mathbf{x}_k = (x_{k1},\ldots,x_{kj},\ldots,x_{kp})$ descritos por p variáveis indexadas por j onde $x_{kj} \in \Re$.

Seja $\mathbf{Y} = \{\mathbf{y}_1,\ldots,\mathbf{y}_c\}$ um conjunto de c protótpos associados a c grupos da partição, tal que cada protótipo de um grupo C_i $(i = 1,\ldots,c)$ é também representado como um vetor de variáveis quantitativas $\mathbf{y}_i = (y_{i1},\ldots,y_{ij},\ldots,y_{ip})$, $y_{ij} \in \Re$. Seja $\mathbb{Y}^c = \underbrace{\mathbb{Y}\times\ldots\times\mathbb{Y}}_{c}$ e $\mathbb{Y} = \underbrace{\Re\times\ldots\times\Re}_{p} = \Re^p$, onde $\mathbb{Y}$ é o espaço de representação dos protótipos tal que $\mathbf{y}_i \in \mathbb{Y}$ e $\mathbf{Y} \in \mathbb{Y}^c$.

Considere $\mathbf{U} = [\mathbf{u}_k]$ uma matriz de matrizes de graus de pertinência multivaridos, onde para cada objeto k de Ω existe uma matriz $c \times p$ de graus de pertinência multivariados $\mathbf{u}_k = [u_{ijk}]$ $(k = 1,\ldots,n)$ e u_{ijk} é o grau de pertinência do objeto k para o grupo C_i $(i = 1,\ldots,c)$ segundo a variável j $(j = 1,\ldots,p)$. Seja $\mathbb{U}^n = \underbrace{\mathbb{U}\times\ldots\times\mathbb{U}}_{n}$, $\mathbb{U} = \underbrace{\mathbb{V}\times\ldots\times\mathbb{V}}_{c}$ e $\mathbb{V} = \underbrace{[0,1]\times\ldots\times[0,1]}_{p}$. Assim, $\mathbb{U}$ é o espaço de representação da matriz de graus de pertinência multivariados $\mathbf{u}_k$ tal que $\mathbf{u}_k \in \mathbb{U}$ e $\mathbf{U} \in \mathbb{U}^n$.

E seja $\Delta = [\delta_{ik}]$ uma matriz $c \times n$ de graus de pertinência padrão onde δ_{ik} é o grau de pertinência do objeto k para o grupo C_i.

3.2 Métodos Propostos

Os métodos apresentados a seguir são baseados na abordagem de agrupamento difuso, onde a principal contribuição é a introdução de graus de pertinência e de possibilidade associados a cada variável (desta forma, passam a ser chamados multivariados). Além disso, também são introduzidas versões poderadas desses métodos.

3.2.1 Método Fuzzy C-Means Multivariado

Embora o método *Fuzzy C-Means* tenha mostrado um bom desempenho em detectar grupos, os valores do grau de pertinência para cada objeto calculado para os grupos não podem indicar quão bem os indivíduos são classificados em relação a cada variável. Isso deve-se ao fato que o grau de pertinência computado neste método leva em conta todas as variáveis, não fazendo distinção entre cada uma. Desta forma, uma abordagem multivariada dos graus de pertinência é proposta.

Considere a Figura 3.1 contendo dois objetos (representados por um losango e um quadrado) e um protótipo (ilustrado por um círculo). Os dois objetos estão a uma mesma distância do protótipo (linha pontilhada). Usando o método FCM para calcular os graus de pertinência, chegaria-se à conclusão de que a chance dos objetos pertecerem ao grupo representado pelo protótipo seria a mesma. De um modo geral, qualquer objeto sobre a linha pontilhada, isto é, equidistante ao protótipo tem o mesmo grau de pertinência. Entretanto, nota-se que o objeto losango está mais distante do protótipo do que o objeto quadrado quanto à variável das abscissas. Por outro lado, essa observação inverte-se para a variável das ordenadas: o objeto losango é mais similiar ao protótipo do que o objeto quadrado. Desta forma, observando cada variável isoladamente é possível obter diferentes similaridades entre objetos e protótipos. Consequentemente, diferentes graus de pertinência para cada variável devem ser calculados a fim de melhor representar a organização dos dados e melhorar a qualidade do agrupamento.

O objetivo do Método *Fuzzy C-Means* Multivariado (aqui abreviado por MFCM) é procurar por uma matriz de protótipos $\mathbf{Y}^*$ e uma matriz de graus de pertinência $\mathbf{U}^*$ tal que:

$$J^1(\mathbf{Y}^*, \mathbf{U}^*) = min\{J^1(\mathbf{Y}, \mathbf{U}) : \mathbf{Y} \in \mathbb{Y}^c, \mathbf{U} \in \mathbb{U}^n\}. \tag{3.1}$$

O algoritmo tem o intuito de minimizar a função objetivo dada como a seguir:

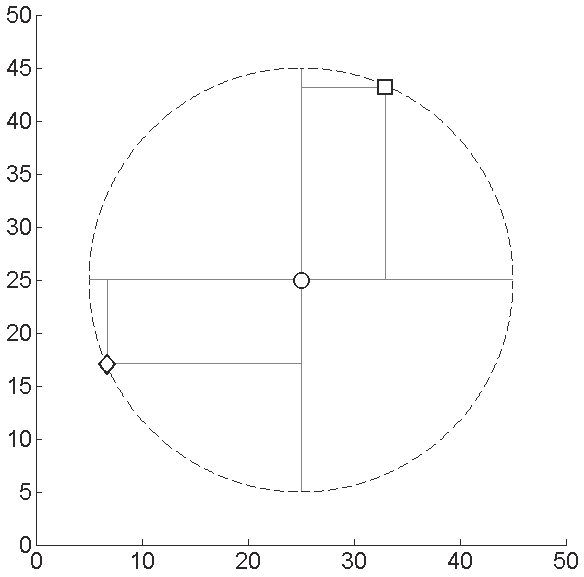

Figura 3.1: Dois objetos equidistantes do protótipo.

$$J^1(\mathbf{Y},\mathbf{U}) = \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk})^m d_{ijk}, \tag{3.2}$$

onde d_{ijk} é a distância Euclidiana que mede a dissimilaridade entre x_{kj} do objeto k e y_{ij} do protótipo do grupo C_i para uma dada variável j. Ela é calculada por:

$$d_{ijk} = \left(x_{kj} - y_{ij}\right)^2. \tag{3.3}$$

Em relação a uma única variável j, os protótipos levam em conta o grau de pertinência e os valores quantitativos de cada objeto de Ω.

Proposição 3.2.1. *Fixando o grau de pertinência u_{ijk}, o protótipo y_{ij} que minimiza o critério J^1 é atualizado usando a seguinte expressão:*

$$y_{ij} = \frac{\sum_{k=1}^{n}(u_{ijk})^m x_{kj}}{\sum_{k=1}^{n}(u_{ijk})^m}. \tag{3.4}$$

Prova 3.2.1. *A prova é dada em A.1.* □

O grau de pertinência multivariado de um dado objeto k de um grupo C_i em relação à variável j é calculado como a seguir.

Proposição 3.2.2. *Fixando o protótipo y_{ij}, o grau de pertinência u_{ijk} que minimiza o critério J^1 é atualizado usando a seguinte equação sob a restrição $\sum_{i=1}^{c}\sum_{j=1}^{p} u_{ijk} = 1$:*

$$u_{ijk} = \left[\sum_{a=1}^{c} \sum_{b=1}^{p} \left(\frac{d_{ijk}}{d_{abk}} \right)^{\frac{1}{m-1}} \right]^{-1}. \tag{3.5}$$

Prova 3.2.2. *A prova é dada em A.2.* □

O grau de pertinência multivariado u_{ijk} de um dado objeto $\mathbf{x}_k$ pertecente ao grupo C_i para uma dada variável j satisfaz as seguintes restrições:

1. $u_{ijk} \in [0,1]$ para todo i, j e k;
2. $0 < \sum_{j=1}^{p} \sum_{k=1}^{n} u_{ijk} < n$ para todo i e
3. $\sum_{i=1}^{c} \sum_{j=1}^{p} u_{ijk} = 1$ para todo k.

No método MFCM, cada objeto tem p graus de pertinência para um dado grupo. Entretanto, ao final da execução do algoritmo, caso o usuário deseje saber a qual grupo cada objeto foi alocado, é necessário encontrar a partição final, isto é, uma partição rígida $P = \{C_1, \ldots, C_i, \ldots, C_c\}$ onde $\bigcup_{i=1}^{c} C_i = \Omega$ e $\bigcap_{i=1}^{c} C_i = \emptyset$. Cada objeto k deve ser alocado ao grupo C_{i*} de forma que:

$$i* = \underbrace{argmax}_{1 \leq i \leq c} \delta_{ik}, \tag{3.6}$$

onde δ_{ik} é o grau de pertinência por grupo de um dado objeto k. A matriz de graus de pertinência por grupo $\Delta = [\delta_{ik}]$ pode ser obtida de diversas formas dependendo como os graus multivariados são agregados. Neste trabalho, a agregação dos graus é feita como a seguir:

$$\delta_{ik} = \sum_{j=1}^{p} u_{ijk}, \tag{3.7}$$

satisfazendo as seguintes condições:

1. $\delta_{ik} \in [0,1]$ para todo i e k;
2. $0 < \sum_{k=1}^{n} \delta_{ik} < n$ para todo i e
3. $\sum_{i=1}^{c} \delta_{ik} = 1$ para todo k.

Algoritmo 7 MFCM (Ω, c)

Requer: Conjunto de dados Ω e o número de grupos c;

1: Faça $\varepsilon > 0$. Fixe c, $2 \leq c \leq n$; fixe m, $1 < m < \infty$; fixe T; e estabeleça $\varepsilon > 0$; Inicialize aleatotiamente $u_{ijk} (i = 1, \ldots, c;\ j = 1, \ldots, p$ e $k = 1, \ldots, n)$ do objeto k pertecente ao grupo C_i para a variável j tal que $u_{ijk} \in [0,1]$, $0 < \sum_{k=1}^{n} u_{ijk} < n$ e $\sum_{i=1}^{c} \sum_{j=1}^{p} u_{ijk} = 1$, para todo $k \in \Omega$.;
2: $t \leftarrow 0$;
3: $J^1(t) \leftarrow 0$;
4: $J^1(t+1) \leftarrow \sum_{i=1}^{c} \sum_{j=1}^{p} \sum_{k=1}^{n} (u_{ijk})^m d_{ijk}$;
5: **Enquanto** $|J^1(t) - J^1(t+1)| > \varepsilon$ **e** $t < T$ **Faça**
6: **Atualize matriz de protótipos Y**: Fixe o grau de pertinência u_{ijk}, atualize os protótipos y_{ij} usando a Equação 3.4;
7: **Atualize matriz de graus de pertinênia U**: Fixe o protótipo y_{ij} $(i = 1, \ldots, c;\ j = 1, \ldots, p$ e $k = 1, \ldots, n)$, atualize o grau de pertinência u_{ijk} usando a Equação 3.5;
8: $J^1(t) \leftarrow J^1(t+1)$;
9: $J^1(t+1) \leftarrow \sum_{i=1}^{c} \sum_{j=1}^{p} \sum_{k=1}^{n} (u_{ijk})^m d_{ijk}$;
10: $t \leftarrow t + 1$;
11: **Fim Enquanto**
12: **Calcule a matriz de grau de pertinência por grupo** Δ: Agregue os graus de pertinência multivariados usando a Equação 3.7.

Retorne Matrizes **Y** e **U**.

O algoritmo do Método *Fuzzy C-Means* Multivariado é descrito como a seguir:

A complexidade de tempo do algoritmo pode ser calculada levando em conta cada uma de suas etapas: Inicialização, Atualização da matriz de protótipos, da matriz de graus de pertinência e do Critério de Parada. Na primeira etapa, o objetivo é inicializar cada valor da matriz de graus de pertinência onde para cada objeto existem cp graus, desta forma serão necessários ncp passos, levando a complexidade ser $O(ncp)$. Na segunda etapa, cada protótipo demanda np processos; como existem c protótipos, então a complexidade será $O(ncp)$. No que se refere ao cálculo dos graus de pertinência, são necessários cp passos para cada objeto, grupo e variável; uma vez que existem n objetos, c grupos e p variáveis, serão necessários ao total nc^2p^2 operações para calcular a matriz de graus de pertinência multivariados, assim a complexidade nesta etapa é $O(nc^2p^2)$. No critério de parada, é necessário calcular o valor da função objetivo em cada iteração do algoritmo, o que demanda ncp operações e levando a complexidade ser $O(ncp)$. Assim, a complexidade de tempo do algoritmo é $O(nc^2p^2)$.

Quanto à complexidade de espaço do algoritmo, o cálculo pode ser feito levando em conta o quanto cada uma das matrizes que o método MFCM demanda. Para armazenar o conjunto de protótipos **Y**, é necessário um espaço p para cada um dos c protótipos, assim a complexidade de espaço é $O(cp)$. Para a matriz de grau de pertinência **U**, existem n matrizes de matrizes de graus de pertinência multivariado, onde cada uma demanda cp, assim o espaço requerido pela matriz **U** é $O(ncp)$. Portanto, a complexidade de espaço do algoritmo MFCM é

$O(ncp)$.

As propriedades de convergência deste tipo de algoritmo podem ser analizadas a partir de duas séries segundo Diday e Simon (1976): $\nu_t = (\mathbf{Y}^t, \mathbf{U}^t) \in \mathbb{Y}^c \times \mathbb{U}^n$ e $\omega_t = J^1(\nu_t) = J^1(\mathbf{Y}^t, \mathbf{U}^t)$, $t = 0, 1, 2, \ldots, T$. O algoritmo começa da série inicial $\nu_0 = (\mathbf{Y}^0, \mathbf{U}^0)$ e calcula os diferentes termos da série ν_t até a convergência quando o critério J^1 atinge um valor estacionário.

Proposição 3.2.3. *A série $\omega_t = J^1(\nu_t)$ decresce a cada iteração e converge.*

Prova 3.2.3. A prova é dada em A.3. □

Proposição 3.2.4. *A série $\nu_t = (\boldsymbol{Y}^t, \boldsymbol{U}^t)$ converge.*

Prova 3.2.4. A prova é dada em A.4. □

3.2.2 Método Fuzzy C-Means Multivariado com Ponderação no Grau de Pertinência

Neste método, cada objeto, grupo e variável tem um peso associado. O objetivo é identificar qual é a relevância de uma determinada variável para calcular o grau de pertinência de um objeto em relação a um grupo. Esta relevância depende da localização do objeto na partição, portanto, cada objeto tem uma interpretação particular, o que pode reduzir a ambiguidade do grau de pertinência. Como exemplo, considere a Figura 3.2. Ela contém dois grupos, onde os protótipos estão na posição $(10, 10)$ e $(40, 10)$, respectivamente. Usando o método MFCM para contruir a matriz de grau de pertinência do objeto (círculo na posição $(24, 11)$), serão obtidos os valores $(0,0335; 0,4686)$ para o Grupo 1 e $(0,0293; 0,4686)$ para o Grupo 2. Isto é, quanto à variável das ordenadas, não existe diferença entre os grupos $(0,4686)$. Por outro lado, embora sejam valores pequenos ($0,0335$ e $0,0293$), os graus de pertinência desse objeto para a variável das abscissas será determinante para diferenciar os grupos. Assim, surge a necessidade de atribuir um peso à variável das abscissas a fim de definir com maior segurança o grupo ao qual o objeto deve ser alocado. O objetivo é atribuir um maior peso ao grau que indicar uma maior possibilidade daquele objeto pertencer a um grupo para uma dada variável.

Seja $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_k]$ uma matriz de matrizes de pesos, onde para um dado objeto k ($k = 1, \ldots, n$), $\boldsymbol{\gamma}_k = [\gamma_{ijk}]$ é uma matriz de pesos $c \times p$ e γ_{ijk} é o peso do grau de pertinência u_{ijk} do objeto k em relação ao grupo C_i ($i = 1, \ldots, c$) e variável j ($j = 1, \ldots, p$).

O objetivo do método *Fuzzy C-Means* Multivariado com ponderação no grau de pertinência (abreviado por MFCM-G) é encontrar um conjunto de protótipos $\mathbf{Y}^*$, uma matriz de matrizes de graus de pertinência multivariado $\mathbf{U}^*$ e uma matriz de matrizes de pesos $\boldsymbol{\Gamma}^*$ tal que:

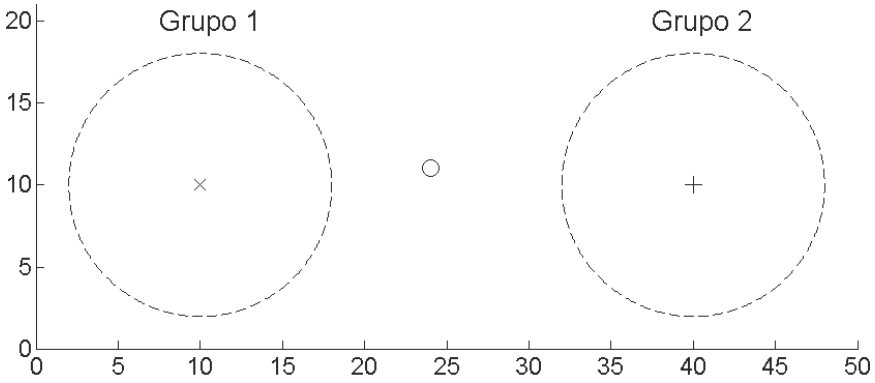

Figura 3.2: Dois protótipos e um objeto em uma região de ambiguidade.

$$J^2(\mathbf{Y}^*,\mathbf{U}^*,\mathbf{\Gamma}^*) = min\{J^2(\mathbf{Y},\mathbf{U},\mathbf{\Gamma}) : \mathbf{Y} \in \mathbb{Y}^c, \mathbf{U} \in \mathbb{U}^n, \mathbf{\Gamma} \in \mathbb{G}^n\}, \tag{3.8}$$

onde $\mathbb{G}^n = \underbrace{\mathbb{G} \times \ldots \times \mathbb{G}}_{n}$, $\mathbb{G} = \underbrace{\mathbb{H} \times \ldots \times \mathbb{H}}_{c}$ e $\mathbb{H} = \underbrace{\Re_+^* \times \ldots \times \Re_+^*}_{p} = {\Re_+^*}^p$, onde $\mathbb{G}$ é o espaço de representação da matriz de pesos γ_k tal que $\gamma_k \in \mathbb{G}$ e $\mathbf{\Gamma} \in \mathbb{G}^n$.

O algoritmo tem como objetivo minimizar a função objetivo dada como a seguir:

$$J^2(\mathbf{Y},\mathbf{U},\mathbf{\Gamma}) = \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}\left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m d_{ijk}. \tag{3.9}$$

Os protótipos são calculados buscando minimizar a função objetivo J^2 e leva em conta os pesos, diferentemente do MFCM apresentado na seção anterior.

Proposição 3.2.5. *Fixo o grau de pertinência u_{ijk} e o peso γ_{ij}, o protótipo y_{ij} do grupo C_i para uma varável j que minimiza o critério J^2 é atualizado usando a seguinte equação:*

$$y_{ij} = \frac{\sum_{k=1}^{n}\left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m x_{kj}}{\sum_{k=1}^{n}\left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m}. \tag{3.10}$$

Prova 3.2.5. A prova pode ser obtida de uma forma similar como descrito em A.1. □

Os graus de pertinência também usam valores dos pesos para representar a chance de um dado objeto pertencer a um grupo para uma dada variável. Mais altos valores do peso implica em maiores graus de pertinência. Assim como o método MFCM, o grau de pertinência multivariado segue as três restrições como definido na seção anterior.

Proposição 3.2.6. *Fixos o peso γ_{ijk} e o protótipo y_{ij}, o grau de pertinência u_{ijk} sob a restrição $\sum_{i=1}^{c}\sum_{j=1}^{p} u_{ijk} = 1$ que minimiza o critério J^2 é atualizado usando a seguinte expressão:*

$$u_{ijk} = \left[\sum_{a=1}^{c}\sum_{b=1}^{p}\left(\frac{d_{ijk}}{d_{abk}}\right)^{\frac{1}{m-1}}\left(\frac{\gamma_{abk}}{\gamma_{ijk}}\right)^{\frac{m}{m-1}}\right]^{-1}. \tag{3.11}$$

Prova 3.2.6. A prova pode ser obtida de uma forma similar como descrito em A.2. □

Por sua vez, o peso no método MFCM-G usa o grau de pertinência e a distância entre um dado objeto e os protótipos a fim de medir a relevância de uma variável para o grau de pertinência final considerando um dado grupo.

Proposição 3.2.7. *Fixos o grau de pertinência u_{ijk} e o protótipo y_{ij}, o peso γ_{ijk} que minimiza o critério J^2 sob a restrição $\sum_{i=1}^{c}\sum_{j=1}^{p}\gamma_{ijk}u_{ijk} = 1$ é atualizado usando a seguinte equação:*

$$\gamma_{ijk} = \frac{\left[(u_{ijk})^m d_{ijk}\right]^{\frac{1}{m+1}}}{\sum_{a=1}^{c}\sum_{b=1}^{p}\left[(u_{abk})^m d_{abk}\right]^{\frac{1}{m+1}} u_{abk}}. \tag{3.12}$$

Prova 3.2.7. A prova é dada em A.5. □

No método MFCM-G, a matriz grau de pertinência final $\Delta = [\delta_{ik}]$ é tal que δ_{ik} é a combinação entre os pesos e os graus de pertinência multivariado, e satisfaz as três condições anteriormente apresentadas. O grau δ_{ik} é dado por:

$$\delta_{ik} = \sum_{j=1}^{p}\gamma_{ijk}u_{ijk}. \tag{3.13}$$

O algoritmo do método *Fuzzy C-Means* Multivariado com ponderação no grau de pertinência é descrito como a seguir:

A complexidade de tempo do algoritmo pode ser semelhantemente entendida como foi descrito para o método MFCM. Ou seja, para a inicialização, a complexidade é $O(ncp)$; na etapa de atualização dos protótipos, é $O(ncp)$; e na etapa de atualização dos graus de pertinência, a complexidade é $O(nc^2p^2)$. Além disso, existe mais uma parte do algoritmo responsável pelo cálculo dos pesos. Para um dado grupo, variável e objeto, são necessários cp operações. Como existem c grupos, p variáveis e n objetos, a construção da matriz de pesos demanda uma complexidade de tempo $O(nc^2p^2)$. O número de operações para o cálculo da função objetivo no critério de parada é igual a ncp, isto é, a complexidade é $O(ncp)$. Finalmente, somando-se as ordens de cada etapa, a complexidade de tempo do algoritmo é $O(nc^2p^2)$.

Algoritmo 8 MFCM-G (Ω, c)

Requer: Conjunto de dados Ω e o número de grupos c;

1: Seja $\varepsilon > 0$. Fixe c, $2 \leq c \leq n$; Fixe m, $1 < m < \infty$; Fixe T; e Fixe $\varepsilon > 0$; Inicialize aleatoriamente $u_{ijk} (i = 1, \ldots, c;\ j = 1, \ldots, p$ e $k = 1, \ldots, n)$ do objeto k pertecente ao grupo C_i para a variável j tal que $u_{ijk} \in [0,1]$, $0 < \sum_{k=1}^{n} u_{ijk} < n$ e $\sum_{i=1}^{c} \sum_{j=1}^{p} u_{ijk} = 1$, para todo $k \in \Omega$. Seja $\gamma_{ijk} = 1$, para todo i, j e k;
2: $t \leftarrow 0$;
3: $J^2(t) \leftarrow 0$;
4: $J^2(t+1) \leftarrow \sum_{i=1}^{c} \sum_{j=1}^{p} \sum_{k=1}^{n} \left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m d_{ijk}$;
5: **Enquanto** $|J^2(t) - J^2(t+1)| > \varepsilon$ **e** $t < T$ **Faça**
6: **Atualize matriz de protótipos Y**: Fixe o grau de pertinência u_{ijk} e peso γ_{ijk} $(i = 1, \ldots, c;\ j = 1, \ldots, p$ e $k = 1, \ldots, n)$ atualize os protótipos y_{ij} usando a Equação 3.10;
7: **Atualize matriz de graus de pertinência U**: Fixe o peso γ_{ijk} e o protótipo y_{ij} $(i = 1, \ldots, c;\ j = 1, \ldots, p$ e $k = 1, \ldots, n)$ atualize o grau de pertinência u_{ijk} usando a Equação 3.11;
8: **Atualize matriz de pesos Γ**: Fixe o grau de pertinência u_{ijk} e o protótipo y_{ij} $(i = 1, \ldots, c;\ j = 1, \ldots, p$ e $k = 1, \ldots, n)$ atualize o peso γ_{ijk} usando a Equação 3.12;
9: $J^2(t) \leftarrow J^2(t+1)$;
10: $J^2(t+1) \leftarrow \sum_{i=1}^{c} \sum_{j=1}^{p} \sum_{k=1}^{n} \left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m d_{ijk}$;
11: $t \leftarrow t + 1$;
12: **Fim Enquanto**
13: **Calcule a matriz de grau de pertinência por grupo** Δ: Agregue os graus de pertinência multivariado usando a Equação 3.13.

Retorne Matrizes **Y**, **U** e **Γ**.

Ao que se refere à complexidade de espaço, existem três principais entidades no algoritmo que demandam armazenamento. O conjunto de protótipos, como visto na seção anterior, requer $O(cp)$. A matriz de graus de pertinência multivariado necessita de $O(ncp)$ de espaço. Enquanto a matriz de pesos demanda mesmo espaço que a matriz de graus de pertinência, isto é, para um dado objeto, exige-se $O(cp)$ e, uma vez que há n ojbetos, então a matriz demanda $O(ncp)$ de espaço. Portanto, a complexidade de espaço para o algoritmo é $O(ncp)$.

Para o algoritmo do método MFCM-G e seguindo Diday e Simon (1976), duas séries pode ser consideradas: $\nu_t = (\mathbf{Y}^t, \mathbf{U}^t, \mathbf{\Gamma}^t) \in \mathbb{Y}^c \times \mathbb{U}^n \times \mathbb{G}^n$ e $\omega_t = J^2(\nu_t) = J^2(\mathbf{Y}^t, \mathbf{U}^t, \mathbf{\Gamma}^t)$, $t = 0, 1, 2, \ldots, T$. O algoritmo começa de uma série inicial $\nu_0 = (\mathbf{Y}^0, \mathbf{U}^0, \mathbf{\Gamma}^0)$ e calcula os diferentes termos das séries ν_t até atingir sua convergência quando o critério J^2 alcança um valor estacionário.

Proposição 3.2.8. *A série $\omega_t = J^2(\nu_t)$ decresce a cada iteração e converge.*

Prova 3.2.8. A prova é dada em A.6. □

Proposição 3.2.9. *A série $\nu_t = (\mathbf{Y}^t, \mathbf{U}^t, \Gamma^t)$ converge.*

Prova 3.2.9. A prova é dada em A.7. □

3.2.3 Método Fuzzy C-Means Multivariado com Ponderação na Distância

Este método apresenta uma distância ponderada a fim de levar em conta a variabilidade de cada variável e grupo, desta forma, os pesos que mudam a cada iteração do algoritmo são computadas. A vantagem destas distâncias ponderadas é que o algoritmo de agrupamento é capaz de encontrar grupos de diferentes formas e tamanhos. Figura 3.3 mostra duas partições contendo dois grupos onde protótipos estão na posição $(10,10)$ e $(30,10)$, respectivamente. Além disso, considere um objeto na posição $(20,11)$. Nota-se que o objeto deve ser alocado ao Grupo 1 na partição do lado esquerdo e Grupo 2 no lado direito. No entanto, o objeto irá ter a mesma matriz de graus de pertinência multivariado se o método MFCM é usado. Assim, os pesos apropriados que representem a forma dos grupos devem ser consideradas com o objetivo de melhorar a qualidade do agrupamento.

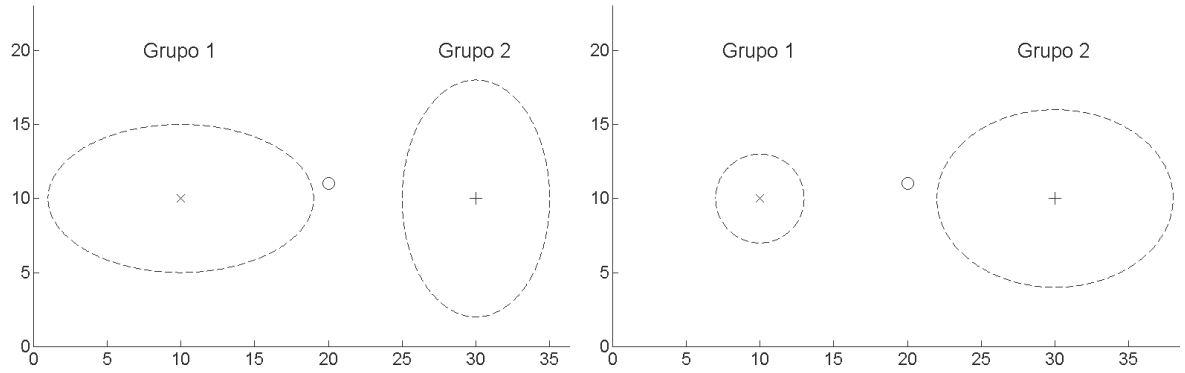

Figura 3.3: Partições contendo grupos de diferentes formas e tamanhos.

Seja $\boldsymbol{\Lambda} = [\boldsymbol{\lambda}_i]$ uma matriz de vetores de pesos, onde $\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{ij}, \ldots, \lambda_{ip})$ e $\lambda_{ij} \in \Re_+^*$ é o peso em relação ao grupo C_i $(i = 1, \ldots, c)$ e variável j $(j = 1, \ldots, p)$.

O objetivo do método *Fuzzy C-Means* Multivariado com ponderação na distância (aqui abrviado como MFCM-D) é encontrar uma matriz de protótipos $\mathbf{Y}^*$, uma matriz de graus de pertinência multivariado $\mathbf{U}^*$ e uma matriz de pesos $\boldsymbol{\Lambda}^*$ tal que:

$$J^3(\mathbf{Y}^*, \mathbf{U}^*, \boldsymbol{\Lambda}^*) = min\{J^3(\mathbf{Y}, \mathbf{U}, \boldsymbol{\Lambda}) : \mathbf{Y} \in \mathbb{Y}^c, \mathbf{U} \in \mathbb{U}^n, \boldsymbol{\Lambda} \in \mathbb{L}^c\}, \tag{3.14}$$

onde $\mathbb{L}^c = \underbrace{\mathbb{L} \times \ldots \times \mathbb{L}}_{c}$, $\mathbb{L} = \underbrace{\Re_+^* \times \ldots \times \Re_+^*}_{p} = \Re_+^{*\,p}$, e $\mathbb{L}$ é o espaço de representação do vetor de pesos $\boldsymbol{\lambda}_i$ tal que $\boldsymbol{\lambda}_i \in \mathbb{L}$ e $\boldsymbol{\Lambda} \in \mathbb{L}^c$.

A função objetivo $J^3(\mathbf{Y},\mathbf{U},\boldsymbol{\lambda})$ do MFCM-D é dada como:

$$J^3(\mathbf{Y},\mathbf{U},\boldsymbol{\Lambda}) = \sum_{i=1}^{c}\sum_{j=1}^{p}\lambda_{ij}\sum_{k=1}^{n}(u_{ijk})^m d_{ijk}, \tag{3.15}$$

onde d_{ijk} é a distância Euclidiana quadrada entre o objeto x_{kj} e o protótipo y_{ij} para uma dada variável j calculada pela Equação 3.3.

Diferentemente do tradicional FCM, os protótipos no método proposto são calculados levando em conta o grau de pertinência de cada variável. Desta forma, protótipos podem melhor representar seus respectivos grupos.

Proposição 3.2.10. *Fixo o grau de pertinência u_{ijk} e o peso λ_{ij}, o protótipo y_{ij} do grupo C_i para a variável j que minimiza o critério J^3 é atualizado usando a equação:*

$$y_{ij} = \frac{\sum_{k=1}^{n}(u_{ijk})^m x_{kj}}{\sum_{k=1}^{n}(u_{ijk})^m}. \tag{3.16}$$

Prova 3.2.10. A prova pode ser obtida de uma forma similar como descrito em A.1. □

Graus de pertinência multivaraido permitem extrair mais informação do conjunto de dados. Isto é, é possível calcular os gaus de pertinência para cada objeto de um grupo de acordo com cada variável. Este tipo de informação aumenta a qualidade do agrupamento, especialmente quando as variáveis possuem dipersões dissimilares. Assim como o método MFCM, o grau de pertinência multivariado no método MFCM-D também segue as três restrições como definido anteriormente.

Proposição 3.2.11. *Fixo o peso λ_{ij} e o protótipo y_{ij}, o grau de pertinência u_{ijk} sob a restrição $\sum_{i=1}^{c}\sum_{j=1}^{p}u_{ijk}=1$ que minimiza o critério J^3 é atualizado usando a seguinte expressão:*

$$u_{ijk} = \left[\sum_{a=1}^{c}\sum_{b=1}^{p}\left(\frac{\lambda_{ij}d_{ijk}}{\lambda_{ab}d_{abk}}\right)^{\frac{1}{m-1}}\right]^{-1}. \tag{3.17}$$

Prova 3.2.11. A prova pode ser obtida de uma forma similar como descrito em A.2. □

A proposta do peso calcular a dispersão das variáveis do conjunto de dados é aumentar a qualidade do agrupamento. Graus de pertinência multivariado dão mais informação ao método de agrupamento, entretanto a forma dos grupos não é levada em conta. Desta forma, pesos são computados para considerar que os grupos podem ser tanto esféricos quanto elípticos. Para isto, pesos são calculados para cada iteração do algoritmo minimizando a função objetivo J^3.

Proposição 3.2.12. *Fixo o grau de pertinência u_{ijk} e o protótipo y_{ij}, o peso λ_{ij} que minimiza o critério J^3 sob a restrição $\lambda_{ij} > 0$ e $\prod_{j=1}^{p} \lambda_{ij} = 1$ é atualizado usando a seguinte equação:*

$$\lambda_{ij} = \frac{\left\{\prod_{h=1}^{p} \left[\sum_{k=1}^{n} (u_{ihk})^m d_{ihk}\right]\right\}^{\frac{1}{p}}}{\sum_{k=1}^{n} (u_{ijk})^m d_{ijk}}. \tag{3.18}$$

Prova 3.2.12. A prova é dada em A.8. □

Depois da convergência do algoritmo, o grau de pertinência δ_{ik} do objeto k para o grupo C_i pode ser similarmente obtido como apresentada pela Equação 3.7 e segue as três restrições anteriormente apresentadas.

O algoritmo do método *Fuzzy C-Means* Multivariado com ponderação na distância é descrito como a seguir:

Algoritmo 9 MFCM-D (Ω, c)

Requer: Conjunto de dados Ω e o número de grupos c;

1: Seja $\varepsilon > 0$. Fixe c, $2 \leq c \leq n$; fixe m, $1 < m < \infty$; fixe T; e fixe $\varepsilon > 0$; Inicialize aleatoriamente $u_{ijk} (i = 1, \ldots, c;\ j = 1, \ldots, p$ e $k = 1, \ldots, n)$ do objeto k pertecente ao grupo C_i para a variável j tal que $u_{ijk} \in [0,1]$, $0 < \sum_{k=1}^{n} u_{ijk} < n$ e $\sum_{i=1}^{c} \sum_{j=1}^{p} u_{ijk} = 1$, para todo $k \in \Omega$. Seja $\lambda_{ij} = 1$, para todo i e j;
2: $t \leftarrow 0$;
3: $J^3(t) \leftarrow 0$;
4: $J^3(t+1) \leftarrow \sum_{i=1}^{c} \sum_{j=1}^{p} \lambda_{ij} \sum_{k=1}^{n} (u_{ijk})^m d_{ijk}$;
5: **Enquanto** $|J^3(t) - J^3(t+1)| > \varepsilon$ **e** $t < T$ **Faça**
6: **Atualize a matriz de protótipos Y**: Fixe o grau de pertinência u_{ijk} e o peso λ_{ij} $(i = 1, \ldots, c;\ j = 1, \ldots, p$ e $k = 1, \ldots, n)$ e atualize os protótipos y_{ij} usando a Equação 3.16;
7: **Atualize a matriz de graus de pertinência U**: Fixe o peso λ_{ij} e o protótipo y_{ij} $(i = 1, \ldots, c;\ j = 1, \ldots, p$ e $k = 1, \ldots, n)$ e atualize o grau de pertinência u_{ijk} usando a Equação 3.17;
8: **Atualize a matriz de pesos Λ**: Fixe o grau de pertinência u_{ijk} e o protótipo y_{ij} $(i = 1, \ldots, c;\ j = 1, \ldots, p$ e $k = 1, \ldots, n)$ e atualize o peso λ_{ij} usando Equação 3.18;
9: $J^3(t) \leftarrow J^3(t+1)$;
10: $J^3(t+1) \leftarrow \sum_{i=1}^{c} \sum_{j=1}^{p} \lambda_{ij} \sum_{k=1}^{n} (u_{ijk})^m d_{ijk}$;
11: $t \leftarrow t + 1$;
12: **Fim Enquanto**
13: **Calcule a matriz de grau de pertinência por grupo** Δ: Agregue os graus de pertinência multivariados usando a Equação 3.7.

Retorne Matrizes **Y**, **U** e Λ.

A fim de calcular complexidade de tempo do algoritmo, ela pode ser entendida de acordo com os seus diferentes passos. Inicialmente, os parâmetros que serão utilizados durante a execução, como ε, m, T, os pesos e os graus de pertinência são definidos. Esta etapa precisa de ncp operações para a matriz **U** e cp operações para a matriz Λ, assim, a complexidade de

tempo para a etapa de inicialização é de $O(ncp)$. O algoritmo executa iterativamente três etapas principais até a convergência. Na primeira, os protótipos são atualizados e o algoritmo demanda np operações para cada um dos c protótipos, portanto, a complexidade de tempo desta estapa é $O(ncp)$. Na segunda etapa, há ncp valores de graus de pertinência onde cada um precisa de cp operações, assim, a complexidade de tempo desta parte do algoritmo é $O(nc^2p^2)$. Na terceira, c vetores de peso são calculadas, onde cada um contém p pesos que precisam de np operações, assim, a complexidade de tempo desta etapa é $O(ncp^2)$. Portanto, no pior caso, a complexidade de tempo computacional do algoritmo é $O(nc^2p^2)$.

Em termos de complexidade de espaço, a matriz de protótipo $\mathbf{Y}$ requer p espaço para cada um dos c protótipos, portanto, é necessário $O(cp)$ espaço para armazenar esta matriz. Para a matriz de graus de pertinência $\mathbf{U}$, existem n matrizes de matrizes de graus de pertinência multivariado, onde cada um requer cp espaço, assim, a complexidade de espaço da matriz $\mathbf{U}$ é $O(ncp)$. A matriz de peso $\mathbf{\Lambda}$ é formada por c vetores de pesos onde cada um requer p espaço, assim, a complexidade de espaço final para armazenar $\mathbf{\Lambda}$ é $O(cp)$. Portanto, a complexidade de espaço deste algoritmo é $O(ncp)$.

De acordo com Diday e Simon (1976), este tipo de algoritmo possui propriedades que podem ser estudadas a partir de duas séries: $v_t = (\mathbf{Y}^t, \mathbf{U}^t, \mathbf{\Lambda}^t) \in \mathbb{Y}^c \times \mathbb{U}^n \times \mathbb{L}^c$ e $\omega_t = J^3(v_t) = J^3(\mathbf{Y}^t, \mathbf{U}^t, \mathbf{\Lambda}^t)$, $t = 0, 1, 2, \ldots, T$. O algoritmo começa da série inicial $v_0 = (\mathbf{Y}^0, \mathbf{U}^0, \mathbf{\Lambda}^0)$ e computa os diferentes termos da série v_t até a convergência quando a função objetivo J^3 chega a um valor estacionário.

Proposição 3.2.13. *A série $\omega_t = J^3(v_t)$ decresce a cada iteração e converge.*

Prova 3.2.13. A prova pode ser obtida de uma forma similar como descrito em A.6. □

Proposição 3.2.14. *A série $v_t = (\mathbf{Y}^t, \mathbf{U}^t, \Lambda^t)$ converge.*

Proof 3.2.14. A prova pode ser obtida de uma forma similar como descrito em A.7. □

3.2.4 Método Possibilistic Fuzzy C-Means Multivariado

Este método diferencia-se dos outros por utilizar a abordagem possibilística. Foi primeiramente proposta para o problema de agrupamento difuso por Krishnapuram & Keller (1993) através do método chamado *Possibilistic C-Means* (PCM) (Krishnapuram e Keller, 1993). A principal vantagem deste método é que ele não é tão sensível a dados aberrantes quanto o FCM. Entretanto, PCM possui a desvantagem de criar grupos coincidentes, isto é, os protótipos

são similares. Para unir os pontos positivos do PCM com o FCM, Pal et al. (2005) propuseram o método *Possibilistic Fuzzy C-Means* (PFCM). Aqui, a proposta é estender este método através do cálculo de graus de pertinência e de possibilidade multivariado com o objetivo de aumentar a quantidade de informação durante o agrupamento, melhorando, assim, a qualidade da partição encontrada.

A Figura 3.4 exemplifica uma configuração onde existem dados aberrante (representandos por círculos). A presença deste tipo de dado pode influenciar no cálculo dos protótipos dos grupos, já que métodos como o *K-Means* ou *Fuzzy C-Means* dão igual importância para este tipo de dados. O método PCM, por sua vez, atribui baixos valores de pertinência para objetos afastados dos grupos de forma que os protótipos representem melhor o conjunto de dados.

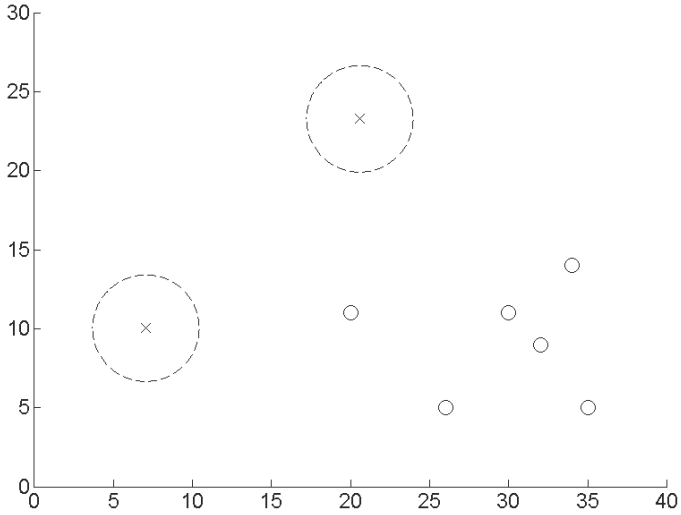

Figura 3.4: Conjunto contendo dados aberrantes.

Seja $\mathbf{T} = [\mathbf{t}_k]$ uma matriz de matrizes de graus de possibilidade multivaridos, onde para um dado objeto k $(k = 1,\ldots,n)$, $\mathbf{t}_k = [t_{ijk}]$ é uma matriz de graus de possibilidade multivaridos $c \times p$ e $t_{ijk} \in [0,1]$ é o grau de possibilidade do objeto k em relação ao grupo C_i $(i = 1,\ldots,c)$ e variável j $(j = 1,\ldots,p)$.

O objetivo do método *Possibilistic Fuzzy C-Means* Multivariado (aqui abreviado como MPFCM) é encontrar uma matriz de protótipos $\mathbf{Y}^*$, uma matriz de graus de pertinência multivariado $\mathbf{U}^*$ e uma matriz de graus de possibilidade multivariado $\mathbf{T}^*$ tal que:

$$J^4(\mathbf{Y}^*,\mathbf{U}^*,\mathbf{T}^*) = min\{J^4(\mathbf{Y},\mathbf{U},\mathbf{T}) : \mathbf{Y} \in \mathbb{Y}^c, \mathbf{U} \in \mathbb{U}^n, \mathbf{T} \in \mathbb{T}^n\}, \tag{3.19}$$

onde $\mathbb{T}^n = \underbrace{\mathbb{T} \times \ldots \times \mathbb{T}}_{n}$, $\mathbb{T} = \underbrace{\mathbb{X} \times \ldots \times \mathbb{X}}_{c}$ e $\mathbb{X} = \underbrace{[0,1] \times \ldots \times [0,1]}_{p}$, onde $\mathbb{T}$ é o espaço de repre-

sentação da matriz de graus de possibilidades multivariados $\mathbf{t}_k$ tal que $\mathbf{t}_k \in \mathbb{T}$ e $\mathbf{T} \in \mathbb{T}^n$.

O algoritmo tem como objetivo minimizar a função objetivo dada como a seguir:

$$J^4(\mathbf{Y},\mathbf{U},\mathbf{T}) = \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m + b(t_{ijk})^\eta\right] d_{ijk} + \pi_{ij}(1-t_{ijk})^\eta\right\}, \tag{3.20}$$

onde a e b são constantes que definem a importânia relativa do grau de pertinência e de possibilidade na função objetivo (Pal et al., 2005). Devem pertencer ao conjunto dos números reais positivos não-nulos, isto é, $a,b \in \Re_*^+$. Ao fazer b maior que a, os protótipos serão influenciados pelos valores dos graus de possibilidade. Desta forma, deve-se usar valores de b maiores que a quando o conjunto de dados contiver *outliers*. Um raciocínio semelhante pode ser aplicado aos parâmetros m e η. Ao fazer $a = 0$, o algoritmo torna-se o *Possibilistic C-Means* (PCM). Por outro lado, ao fazer $b = 0$ e $\pi_{ij} = 0$, o algoritmo torna-se o *Fuzzy C-Means* (FCM).

O valor de π_{ij} tem a importância de balancear os dois termos da equação. A ideia é que π_{ij} tenha a mesma ordem de grandeza de d_{ijk}. Assim, uma proposta dada pelos autores Krishnapuram e Keller para calcular π_i no PCM e que também foi adotado por Pal et al. no PFCM é usar a média das distâncias objeto-protótipo. Aqui, a principal diferença é que este valor é calculado para cada variável j, levando em conta o comportamento de cada uma. É dado por:

$$\pi_{ij} = K\frac{\sum_{k=1}^{n}(u_{ijk})^m d_{ijk}}{\sum_{k=1}^{n}(u_{ijk})^m}, \tag{3.21}$$

onde K é frequentemente utilizado como 1. No algoritmo, π_{ij} é calculado uma única vez logo após a inicialização dos parâmetros para evitar instabilidades na convergência.

Os protótipos no método MPFCM levam em conta os valores a, b e os graus multivariados: de pertinência u_{ijk} e de possibilidade t_{ijk}. Desta forma, estes valores exercem uma grande influência na representação dos grupos e consequentemente na qualidade do agrupamento.

Proposição 3.2.15. *Fixo o grau de pertinência u_{ijk} e o grau de possibilidade t_{ijk}, o protótipo y_{ij} que minimiza o critério J^4 é atualizado usando a seguinte expressão:*

$$y_{ij} = \frac{\sum_{k=1}^{n}\left[a(u_{ijk})^m + b(t_{ijk})^\eta\right] x_{kj}}{\sum_{k=1}^{n}\left[a(u_{ijk})^m + b(t_{ijk})^\eta\right]} \tag{3.22}$$

Prova 3.2.15. A prova pode ser obtida de uma forma similar como descrito em A.1. □

O grau de pertinência herda as mesmas propriedades dos métodos da aboradagem difusa multivariada apresentados nas seções anteriores (MFCM, MFCM-G, MFCM-D): está sob a

restrição do somatório e é utilizado para indicar a pertinência de um objeto para um dado grupo em relação a uma variável.

Proposição 3.2.16. *Fixo o protótipo y_{ij} e o grau de possibilidade t_{ijk}, o grau de pertinência u_{ijk} sob a restrição $\sum_{i=1}^{c}\sum_{j=1}^{p} u_{ijk} = 1$ que minimiza o critério J^4 é atualizado usando a seguinte expressão:*

$$u_{ijk} = \left[\sum_{a=1}^{c}\sum_{b=1}^{p}\left(\frac{d_{ijk}}{d_{abk}}\right)^{\frac{1}{m-1}}\right]^{-1}. \tag{3.23}$$

Prova 3.2.16. A prova pode ser obtida de uma forma similar como descrito em A.2. □

O grau de possibilidade diferencia-se do grau de pertinência pelo fato que o primeiro não precisa estar sob a restição da soma, isto é, $\sum_{i=1}^{c}\sum_{j=1}^{p} t_{ijk}$ não seja necessariamente igual a 1. Essa flexibilidade permite que os métodos baseados na abordagem possibilística sejam mais robustos a dados aberrantes. Quanto mais afastado um objeto estiver do protótipo, menor será o grau de possibilidade de pertencer ao grupo, fazendo com que este objeto exerça menos influência no cálculo dos representantes e consequentemente aumentando a qualidade do agrupamento.

Proposição 3.2.17. *Fixo o protótipo y_{ij} e o grau de pertinência u_{ijk}, o grau de possibilidade t_{ijk} que minimiza o critério J^4 é atualizado usando a seguinte expressão:*

$$t_{ijk} = \left[1 + \left(b\frac{d_{ijk}}{\pi_{ij}}\right)^{\frac{1}{(\eta-1)}}\right]^{-1}. \tag{3.24}$$

Prova 3.2.17. A prova é dada em A.9. □

O algoritmo do método *Possibilistic Fuzzy C-Means* Multivariado é descrito como a seguir:

Com o propósito de calcular a complexidade de tempo do algoritmo, ela pode ser entendida de acordo com os seus diferentes passos. Na inicialização, esta etapa precisa de ncp operações para a matriz **U** e também ncp operações para a matriz **T**, assim, a complexidade de tempo para esta primeira estapa é de $O(ncp)$. Equanto a convergência não for obtida, o algoritmo executa iterativamente três etapas principais. Na primeira, os protótipos são atualizados e o algoritmo requer np operações para cada um dos c protótipos, portanto, a complexidade de tempo desta etapa é $O(ncp)$. Na segunda etapa, há ncp valores de graus de pertinência onde cada um precisa de cp operações, assim, a complexidade de tempo desta parte do algoritmo é $O(nc^2p^2)$. Na terceira, n matrizes de graus de possibilidades são calculadas, onde cada uma

Algoritmo 10 MPFCM (Ω, c)

Requer: Conjunto de dados Ω e o número de grupos c;

1: Seja $\varepsilon > 0$. Fixe c, $2 \leq c \leq n$; fixe $a > 0$ e $b > 0$; fixe m, $1 < m < \infty$; fixe η, $1 < \eta < \infty$; fixe T; e fixe $\varepsilon > 0$; Inicialize aleatoriamente u_{ijk} e t_{ijk} $(i = 1,\ldots,c;\ j = 1,\ldots,p$ e $k = 1,\ldots,n)$ do objeto k pertecente ao grupo C_i para a variável j tal que $u_{ijk} \in [0,1]$, $0 < \sum_{k=1}^{n} u_{ijk} < n$, $\sum_{i=1}^{c}\sum_{j=1}^{p} u_{ijk} = 1$ e $t_{ijk} > 0$, para todo $k \in \Omega$. Calcule o valor de π_{ij} $(i = 1,\ldots,c$ e $j = 1,\ldots,p)$ usando a Equação 3.21.
2: $t \leftarrow 0$;
3: $J^4(t) \leftarrow 0$;
4: $J^4(t+1) \leftarrow \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m + b(t_{ijk})^\eta\right] d_{ijk} + \pi_{ij}(1-t_{ijk})^\eta\right\}$;
5: **Enquanto** $|J^4(t) - J^4(t+1)| > \varepsilon$ **e** $t < T$ **Faça**
6: **Atualize a matriz de protótipos Y**: Fixe o grau de pertinência u_{ijk} e o grau de possibilidade t_{ijk} $(i = 1,\ldots,c;\ j = 1,\ldots,p$ e $k = 1,\ldots,n)$ e atualize os protótipos y_{ij} usando a Equação 3.22;
7: **Atualize a matriz de graus de pertinência U**: Fixe o grau de possibilidade t_{ijk} e o protótipo y_{ij} $(i = 1,\ldots,c;\ j = 1,\ldots,p$ e $k = 1,\ldots,n)$ e atualize o grau de pertinência u_{ijk} usando a Equação 3.23;
8: **Atualize a matriz de graus de possibilidade T**: Fixe o grau de pertinência u_{ijk} e o protótipo y_{ij} $(i = 1,\ldots,c;\ j = 1,\ldots,p$ e $k = 1,\ldots,n)$ e atualize o grau de possibilidade t_{ijk} usando Equação 3.24;
9: $J^4(t) \leftarrow J^4(t+1)$;
10: $J^4(t+1) \leftarrow \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m + b(t_{ijk})^\eta\right] d_{ijk} + \pi_{ij}(1-t_{ijk})^\eta\right\}$;
11: $t \leftarrow t+1$;
12: **Fim Enquanto**
13: **Calcule a matriz de grau de pertinência por grupo** Δ: Agregue os graus de pertinência multivariados usando a Equação 3.7.

Retorne Matrizes **Y**, **U** e **T**.

contém cp velores que requerem $O(1)$ operações, assim, a complexidade de tempo desta etapa é $O(ncp)$. Portanto, no pior caso, a complexidade de tempo computacional do algoritmo é $O(nc^2p^2)$.

No que diz respeito à complexidade de espaço, a matriz de protótipo **Y** necessita de p espaço para cada um dos c protótipos, portanto, é necessário $O(cp)$ espaço para armazenar esta matriz. Para a matriz de graus de pertinência **U**, existem n matrizes de matriz de graus de pertinência multivariado, onde cada um requer cp espaço, assim, a complexidade de espaço da matriz **U** é $O(ncp)$. A matriz de graus de possibilidades multivariado **T** segue o mesmo raciocínio da matriz **U** e desta forma requer $O(ncp)$ espaço. Portanto, a complexidade de espaço deste algoritmo é $O(ncp)$.

Segundo Diday e Simon (1976), o algoritmo possui propriedades que podem ser estudadas a partir de duas séries: $\nu_t = (\mathbf{Y}^t, \mathbf{U}^t, \mathbf{T}^t) \in \mathbb{Y}^c \times \mathbb{U}^n \times \mathbb{T}^n$ e $\omega_t = J^4(\nu_t) = J^4(\mathbf{Y}^t, \mathbf{U}^t, \mathbf{T}^t)$, $t = 0,1,2,\ldots,T$. O algoritmo começa da série inicial $\nu_0 = (\mathbf{Y}^0, \mathbf{U}^0, \mathbf{T}^0)$ e computa os diferentes termos da série ν_t até a convergência quando a função objetivo J^4 chega a um valor estacionário.

Proposição 3.2.18. *A série $\omega_t = J^4(\nu_t)$ decresce a cada iteração e converge.*

Prova 3.2.18. A prova pode ser obtida de uma forma similar como descrito em A.6. □

Proposição 3.2.19. *A série* $v_t = (\boldsymbol{Y}^t, \boldsymbol{U}^t, \boldsymbol{T}^t)$ *converge.*

Proof 3.2.19. A prova pode ser obtida de uma forma similar como descrito em A.7. □

3.2.5 Método Possibilistic Fuzzy C-Means Multivariado Ponderado

A fim de levar em conta a variabilidade de cada variável e grupo, este método apresenta distâncias ponderadas, desta forma, os pesos que mudam a cada iteração do algoritmo são computadas. Com este tipo de ponderação, o método torna-se uma generalização da sua versão sem peso como apresentado na seção anterior, uma vez que distâncias ponderadas tornam os algoritmos capazes de encontrar grupos de diferentes formas e tamanhos. O objetivo é utilizar pesos como uma forma de quantificação da importância de cada variável em um determinado grupo levando em conta a sua variabilidade, isto é, a dispersão dos objetos em relação o protótipo. Figura 3.5 mostra uma partição contendo três grupos de diferentes formas, onde os círculos representam dados aberrantes. Nota-se que, em relação ao Grupo 3, embora os dados aberrantes tanto na parte superior quanto na inferior estejam aproximadamente equidistantes do centro deste grupo, a interpretação desse tipo de dados pode variar se levada em conta a dispersão das variáveis. Além disso, similarmente ao método MFCM-D, a distância ponderada pode melhorar a qualidade do agrupamento por considerar uma informação a mais a respeito da partição: a variabilidade das variáveis em cada grupo.

O objetivo do método *Possibilistic Fuzzy C-Means* Multivariado Poderado (aqui abreviado como MPFCM-D) é encontrar uma matriz de protótipos $\mathbf{Y}^*$, uma matriz de graus de pertinência multivariado $\mathbf{U}^*$, uma matriz de graus de possibilidade multivariado $\mathbf{T}^*$ e matriz de peso $\boldsymbol{\Lambda}$ tal que:

$$J^5(\mathbf{Y}^*, \mathbf{U}^*, \mathbf{T}^*, \boldsymbol{\Lambda}^*) = min\{J^5(\mathbf{Y}, \mathbf{U}, \mathbf{T}, \boldsymbol{\Lambda}) : \mathbf{Y} \in \mathbb{Y}^c, \mathbf{U} \in \mathbb{U}^n, \mathbf{T} \in \mathbb{T}^n, \boldsymbol{\Lambda} \in \mathbb{L}^c\}, \tag{3.25}$$

onde $\mathbb{L}^c = \underbrace{\mathbb{L} \times \ldots \times \mathbb{L}}_{c}$, $\mathbb{L} = \underbrace{\Re_+^* \times \ldots \times \Re_+^*}_{p} = \Re_+^{*\,p}$, e $\mathbb{L}$ é o espaço de representação do vetor de pesos $\boldsymbol{\lambda}_i$ tal que $\boldsymbol{\lambda}_i \in \mathbb{L}$ e $\boldsymbol{\Lambda} \in \mathbb{L}^c$.

O algoritmo tem como objetivo minimizar a função objetivo dada como a seguir:

$$J^5(\mathbf{Y}, \mathbf{U}, \mathbf{T}, \boldsymbol{\Lambda}) = \sum_{i=1}^{c} \sum_{j=1}^{p} \lambda_{ij} \sum_{k=1}^{n} \left\{ \left[a(u_{ijk})^m + b(t_{ijk})^\eta \right] d_{ijk} + \pi_{ij}(1 - t_{ijk})^\eta \right\}, \tag{3.26}$$

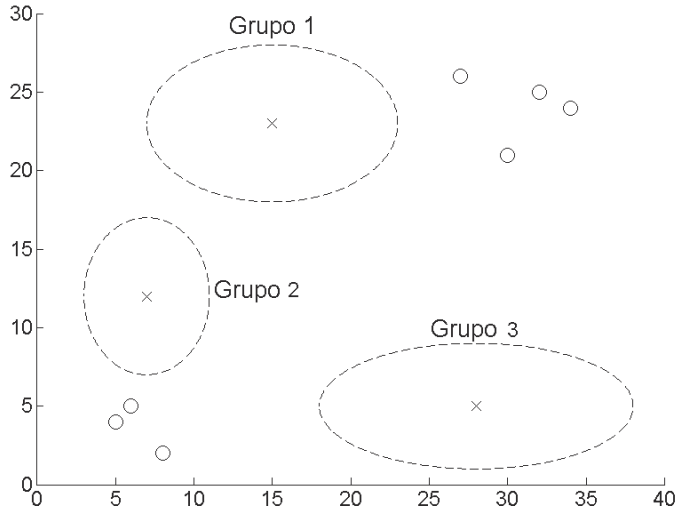

Figura 3.5: Conjunto contendo dados aberrantes e grupos de diferentes formas e tamanhos.

O valor de π_{ij} segue a Equação 3.21 como descrita na seção anterior.

Proposição 3.2.20. *Fixo o grau de pertinência u_{ijk}, o grau de possibilidade t_{ijk} e o peso λ_{ij}, o protótipo y_{ij} que minimiza o critério J^5 é atualizado usando a seguinte expressão:*

$$y_{ij} = \frac{\sum_{k=1}^{n} \left[a(u_{ijk})^m + b(t_{ijk})^\eta \right] x_{kj}}{\sum_{k=1}^{n} \left[a(u_{ijk})^m + b(t_{ijk})^\eta \right]} \tag{3.27}$$

Prova 3.2.20. A prova pode ser obtida de uma forma similar como descrito em A.1. □

O grau de pertinência é similarmente calculado como nos métodos anteriormente descritos. Nesse método, leva-se em conta o peso λ_{ij}.

Proposição 3.2.21. *Fixo o protótipo y_{ij}, o grau de possibilidade t_{ijk} e o peso λ_{ij}, o grau de pertinência u_{ijk} que minimiza o critério J^5 é atualizado usando a seguinte expressão:*

$$u_{ijk} = \left[\sum_{a=1}^{c} \sum_{b=1}^{p} \left(\frac{\lambda_{ij} d_{ijk}}{\lambda_{ab} d_{abk}} \right)^{\frac{1}{m-1}} \right]^{-1}. \tag{3.28}$$

Prova 3.2.21. A prova pode ser obtida de uma forma similar como descrito em A.2. □

Neste método, o grau de possibilidade é similarmente computado como no método anterior.

Proposição 3.2.22. *Fixo o protótipo y_{ij}, o grau de pertinência u_{ijk} e o peso λ_{ij}, o grau de possibilidade t_{ijk} que minimiza o critério J^5 é atualizado usando a seguinte expressão:*

$$t_{ijk} = \left[1 + \left(b\frac{\lambda_{ij} d_{ijk}}{\pi_{ij}}\right)^{\frac{1}{(\eta-1)}}\right]^{-1}. \tag{3.29}$$

Prova 3.2.22. A prova pode ser obtida de uma forma similar como descrito em A.9. □

Proposição 3.2.23. *Fixo o protótipo y_{ij}, o grau de pertinência u_{ijk} e o grau de possibilidade t_{ijk}, o valor de λ_{ij} sob a restrição $\prod_{h=1}^{p} \lambda_{ih} = 1$ que minimiza o critério J^5 é atualizado usando a seguinte expressão:*

$$\lambda_{ij} = \frac{\left\{\prod_{h=1}^{p}\left[\sum_{k=1}^{n}\left\{\left[a(u_{ihk})^m + b(t_{ihk})^\eta\right] d_{ihk} + \pi_{ih}(1-t_{ihk})^\eta\right\}\right]^{\frac{1}{p}}\right\}}{\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m + b(t_{ijk})^\eta\right] d_{ijk} + \pi_{ij}(1-t_{ijk})^\eta\right\}}. \tag{3.30}$$

Prova 3.2.23. A prova é dada em A.10. □

O algoritmo do método *Possibilistic Fuzzy C-Means* Multivariado Ponderado é descrito como a seguir:

A complexidade de tempo do algoritmo pode ser compreendida semelhantemente como apresentada no método anterior. Na inicialização, a principal diferença é a atribuição dos pesos para valores iguais a 1 o que requer $O(cp)$ para cada matriz, não alterando a complexidade da incialização que é $O(ncp)$. Neste algortimo, enquanto a convergência não é atingida, é realizado a execução de cinco etapas. As três primeiras são similares às do método MPFCM, onde as complexidades são, respectivamente $O(ncp)$, $O(nc^2p^2)$ e $O(ncp)$. Na última etapa existem cp valores de pesos onde cada um requer np operações, assim o tempo necessário para o cálculo da matriz é da ordem de $O(ncp^2)$. Portanto, no pior caso, a complexidade de tempo computacional do algoritmo é $O(nc^2p^2)$.

Referente à complexidade de espaço, a matriz de protótipo $\mathbf{Y}$, de graus de pertinência $\mathbf{U}$ e de graus de possibilidades multivariado $\mathbf{T}$ requerem, respectivamente $O(cp)$, $O(ncp)$ e $O(ncp)$. A matriz de peso $\boldsymbol{\Lambda}$ é formada por c vetores de pesos onde cada um requer p espaço, assim, a complexidade de espaço final para armazenar $\boldsymbol{\Lambda}$ é $O(cp)$. Portanto, a complexidade de espaço do algoritmo é $O(ncp)$.

Segundo Diday e Simon (1976), o algoritmo possui propriedades que podem ser estudadas a partir de duas séries: $\nu_t = (\mathbf{Y}^t, \mathbf{U}^t, \mathbf{T}^t, \boldsymbol{\Lambda}^t) \in \mathbb{Y}^c \times \mathbb{U}^n \times \mathbb{T}^n \times \mathbb{L}^c$ e $\omega_t = J^5(\nu_t) = J^5(\mathbf{Y}^t, \mathbf{U}^t, \mathbf{T}^t, \mathbf{L}^t)$, $t = 0, 1, 2, \ldots, T$. O algoritmo começa da série inicial $\nu_0 = (\mathbf{Y}^0, \mathbf{U}^0, \mathbf{T}^0, \boldsymbol{\Lambda}^0)$

Algoritmo 11 MPFCM-D (Ω, c)

Requer: Conjunto de dados Ω e o número de grupos c;

1: Seja $\varepsilon > 0$. Fixe c, $2 \leq c \leq n$; fixe $a > 0$ e $b > 0$; fixe m, $1 < m < \infty$; fixe η, $1 < \eta < \infty$; fixe T; e fixe $\varepsilon > 0$; Inicialize aleatoriamente u_{ijk} e t_{ijk} $(i = 1, \ldots, c;\ j = 1, \ldots, p$ e $k = 1, \ldots, n)$ do objeto k pertecente ao grupo C_i para a variável j tal que $u_{ijk} \in [0,1]$, $0 < \sum_{k=1}^{n} u_{ijk} < n$, $\sum_{i=1}^{c} \sum_{j=1}^{p} u_{ijk} = 1$ e $t_{ijk} > 0$, para todo $k \in \Omega$. Calcule o valor de π_{ij} $(i = 1, \ldots, c$ e $j = 1, \ldots, p)$ usando a Equação 3.21. Inicialize o peso $\lambda_{ij} = 1$ $(i = 1, \ldots, c$ e $j = 1, \ldots, p)$.

2: $t \leftarrow 0$;

3: $J^5(t) \leftarrow 0$;

4: $J^5(t+1) \leftarrow \sum_{i=1}^{c} \sum_{j=1}^{p} \lambda_{ij} \sum_{k=1}^{n} \left\{ \left[a(u_{ijk})^m + b(t_{ijk})^\eta \right] d_{ijk} + \pi_{ij}(1 - t_{ijk})^\eta \right\}$;

5: **Enquanto** $|J^5(t) - J^5(t+1)| > \varepsilon$ **e** $t < T$ **Faça**

6: **Atualize a matriz de protótipos Y**: Fixe o grau de pertinência u_{ijk}, o grau de possibilidade t_{ijk} e o peso λ_{ij} $(i = 1, \ldots, c;\ j = 1, \ldots, p$ e $k = 1, \ldots, n)$ e atualize os protótipos y_{ij} usando a Equação 3.27;

7: **Atualize a matriz de graus de pertinência U**: Fixe o grau de possibilidade t_{ijk}, o protótipo y_{ij} e o peso λ_{ij} $(i = 1, \ldots, c;\ j = 1, \ldots, p$ e $k = 1, \ldots, n)$ e atualize o grau de pertinência u_{ijk} usando a Equação 3.28;

8: **Atualize a matriz de graus de possibilidade T**: Fixe o grau de pertinência u_{ijk}, o protótipo y_{ij} e o peso λ_{ij} $(i = 1, \ldots, c;\ j = 1, \ldots, p$ e $k = 1, \ldots, n)$ e atualize o grau de possibilidade t_{ijk} usando Equação 3.29;

9: **Atualize a matriz de pesos Λ**: Fixe o grau de pertinência u_{ijk}, o grau de possibilidade t_{ijk} e o protótipo y_{ij} $(i = 1, \ldots, c;\ j = 1, \ldots, p$ e $k = 1, \ldots, n)$ e atualize o peso λ_{ij} usando Equação 3.30;

10: $J^5(t) \leftarrow J^5(t+1)$;

11: $J^5(t+1) \leftarrow \sum_{i=1}^{c} \sum_{j=1}^{p} \lambda_{ij} \sum_{k=1}^{n} \left\{ \left[a(u_{ijk})^m + b(t_{ijk})^\eta \right] d_{ijk} + \pi_{ij}(1 - t_{ijk})^\eta \right\}$;

12: $t \leftarrow t + 1$;

13: **Fim Enquanto**

14: **Calcule a matriz de grau de pertinência por grupo** Δ: Agregue os graus de pertinência multivariados usando a Equação 3.7.

Retorne Matrizes **Y**, **U**, **T** e Λ.

e computa os diferentes termos da série v_t até a convergência quando a função objetivo J^5 chega a um valor estacionário.

Proposição 3.2.24. *A série $\omega_t = J^5(v_t)$ decresce a cada iteração e converge.*

Prova 3.2.24. A prova pode ser obtida de uma forma similar como descrito em A.6. □

Proposição 3.2.25. *A série $v_t = (\boldsymbol{Y}^t, \boldsymbol{U}^t, \boldsymbol{T}^t, \Lambda^t)$ converge.*

Proof 3.2.25. A prova pode ser obtida de uma forma similar como descrito em A.7. □

3.3 Interpretação de Partição e Grupo

Índices de interpretação são ferramentas úteis que permitem que o usuário entenda melhor o agrupamento resultante. Celeux et al. (1989) introduziram uma família de índices para a interpretação de grupo que são baseados na soma dos quadrados (SSQ) para dados quantitativos particionado pelos algoritmos de agrupamento rígidos. Nesta seção, adaptações destes índices foram feitas para o caso de dados quantitativos particionado por algoritmos de agrupamento difuso multivariado.

Para as próximos equações, o índice $l (l = 1,2,3,4,5)$ corresponde a um método, onde $l = 1$ está associada ao método MFCM, $l = 2$ para MFCM-G, $l = 3$ para MFCM-D, $l = 4$ para MPFCM e $l = 5$ para MPFCM-D.

3.3.1 Medidas baseadas em Soma de Quadrados

Nesta seção, é definido o SSQ global e o SSQ para intra e inter grupos para métodos difusos multivariados. Soma dos quadrados pode ser decomposta em soma de quadrados por grupo e variável. Além disso, essas decomposições são fundamentais para a definição de índices de interpretação.

A dispersão global de todos os n objetos é calculada de acordo com a função de distância dada pela Equação 3.3. É como se segue:

$$
\begin{aligned}
T^1 &= \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk})^m(x_{kj}-z_j^1)^2 \\
T^2 &= \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}\left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m(x_{kj}-z_j^2)^2 \\
T^3 &= \sum_{i=1}^{c}\sum_{j=1}^{p}\lambda_{ij}\sum_{k=1}^{n}(u_{ijk})^m(x_{kj}-z_j^3)^2 \\
T^4 &= \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](x_{kj}-z_j^4)^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \\
T^5 &= \sum_{i=1}^{c}\sum_{j=1}^{p}\lambda_{ij}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](x_{kj}-z_j^5)^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \qquad (3.31)
\end{aligned}
$$

Para todos os métodos, T mede a dispersão global de todos os objetos, ou seja, mede quão dispersos os objetos estão em relação ao centróide global $\mathbf{z}^l$.

Proposição 3.3.1. *O vetor do centróide global* $\mathbf{z}^l = (z_1^l, \ldots, z_j^l, \ldots, z_p^l)$ *para todos os n objetos, que minimiza a dispersão global* T^l*, tem seus componentes* z_j^l *atualizado por:*

$$z_j^1 = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ijk})^m x_{kj}}{\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ijk})^m}. \tag{3.32}$$

$$z_j^2 = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n}\left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m x_{kj}}{\sum_{i=1}^{c}\sum_{k=1}^{n}\left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m}. \tag{3.33}$$

$$z_j^3 = \frac{\sum_{i=1}^{c}\lambda_{ij}\sum_{k=1}^{n}(u_{ijk})^m x_{kj}}{\sum_{i=1}^{c}\lambda_{ij}\sum_{k=1}^{n}(u_{ijk})^m}. \tag{3.34}$$

$$z_j^4 = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n}\left[a(u_{ijk})^m + b(t_{ijk})^\eta\right] x_{kj}}{\sum_{i=1}^{c}\sum_{k=1}^{n}\left[a(u_{ijk})^m + b(t_{ijk})^\eta\right]} \tag{3.35}$$

$$z_j^5 = \frac{\sum_{i=1}^{c}\lambda_{ij}\sum_{k=1}^{n}\left[a(u_{ijk})^m + b(t_{ijk})^\eta\right] x_{kj}}{\sum_{i=1}^{c}\lambda_{ij}\sum_{k=1}^{n}\left[a(u_{ijk})^m + b(t_{ijk})^\eta\right]} \tag{3.36}$$

Prova 3.3.1. A prova pode ser obtida de uma forma similar como descrito em A.1. □

Uma vez que T^l é aditivo, ele pode ser decomposto na dispersão global T_i^l para o grupo C_i dado por $T^l = \sum_{i=1}^{c} T_i^l (l = 1,2,3,4,5)$ com:

$$\begin{aligned}
T_i^1 &= \sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk})^m (x_{kj} - z_j^1)^2 \\
T_i^2 &= \sum_{j=1}^{p}\sum_{k=1}^{n}\left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m (x_{kj} - z_j^2)^2 \\
T_i^3 &= \sum_{j=1}^{p}\lambda_{ij}\sum_{k=1}^{n}(u_{ijk})^m (x_{kj} - z_j^3)^2 \\
T_i^4 &= \sum_{j=1}^{p}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m + b(t_{ijk})^\eta\right](x_{kj} - z_j^4)^2 + \pi_{ij}(1 - t_{ijk})^\eta\right\} \\
T_i^5 &= \sum_{j=1}^{p}\lambda_{ij}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m + b(t_{ijk})^\eta\right](x_{kj} - z_j^5)^2 + \pi_{ij}(1 - t_{ijk})^\eta\right\}
\end{aligned} \tag{3.37}$$

Um raciocínio semelhante pode ser aplicado ao caso em que a dispersão global é calculada de acordo com a variável j:

$$
\begin{aligned}
T_j^1 &= \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ijk})^m(x_{kj}-z_j^1)^2 \\
T_j^2 &= \sum_{i=1}^{c}\sum_{k=1}^{n}\left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m(x_{kj}-z_j^2)^2 \\
T_j^3 &= \sum_{i=1}^{c}\lambda_{ij}\sum_{k=1}^{n}(u_{ijk})^m(x_{kj}-z_j^3)^2 \\
T_j^4 &= \sum_{i=1}^{c}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](x_{kj}-z_j^4)^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \\
T_j^5 &= \sum_{i=1}^{c}\lambda_{ij}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](x_{kj}-z_j^5)^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \qquad (3.38)
\end{aligned}
$$

Em relação ao grupo C_i e variável j, a disperção global para os métodos tal que $T^l = \sum_{i=1}^{c} T_i^l = \sum_{j=1}^{p}(\sum_{i=1}^{c} T_{ij}^l)$ é dado como:

$$
\begin{aligned}
T_{ij}^1 &= \sum_{k=1}^{n}(u_{ijk})^m(x_{kj}-z_j^1)^2 \\
T_{ij}^2 &= \sum_{k=1}^{n}\left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m(x_{kj}-z_j^2)^2 \\
T_{ij}^3 &= \lambda_{ij}\sum_{k=1}^{n}(u_{ijk})^m(x_{kj}-z_j^3)^2 \\
T_{ij}^4 &= \sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](x_{kj}-z_j^4)^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \\
T_{ij}^5 &= \lambda_{ij}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](x_{kj}-z_j^5)^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \qquad (3.39)
\end{aligned}
$$

A dispersão global intra-grupo mede a soma das diferenças entre cada objeto e protótipos. Quanto mais semelhantes os objetos são dos protótipos, mais concisos são os grupos. Muitos métodos de agrupamento visam reduzir esta medida. Ela é dada como:

$$
\begin{aligned}
J^1 &= \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk})^m(x_{kj}-y_{ij})^2 \\
J^2 &= \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}\left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m(x_{kj}-y_{ij})^2 \\
J^3 &= \sum_{i=1}^{c}\sum_{j=1}^{p}\lambda_{ij}\sum_{k=1}^{n}(u_{ijk})^m(x_{kj}-y_{ij})^2 \\
J^4 &= \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](x_{kj}-y_{ij})^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \\
J^5 &= \sum_{i=1}^{c}\sum_{j=1}^{p}\lambda_{ij}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](x_{kj}-y_{ij})^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \qquad (3.40)
\end{aligned}
$$

A dispersão intra-grupo do grupo C_i é calculadda como a seguir:

$$
\begin{aligned}
J_i^1 &= \sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk})^m(x_{kj}-y_{ij})^2 \\
J_i^2 &= \sum_{j=1}^{p}\sum_{k=1}^{n}\left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m(x_{kj}-y_{ij})^2 \\
J_i^3 &= \sum_{j=1}^{p}\lambda_{ij}\sum_{k=1}^{n}(u_{ijk})^m(x_{kj}-y_{ij})^2 \\
J_i^4 &= \sum_{j=1}^{p}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](x_{kj}-y_{ij})^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \\
J_i^5 &= \sum_{j=1}^{p}\lambda_{ij}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](x_{kj}-y_{ij})^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \qquad (3.41)
\end{aligned}
$$

Em relação à variável j, a dispersão intra-grupo é dada como:

$$
\begin{aligned}
J_j^1 &= \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ijk})^m(x_{kj}-y_{ij})^2 \\
J_j^2 &= \sum_{i=1}^{c}\sum_{k=1}^{n}\left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m(x_{kj}-y_{ij})^2 \\
J_j^3 &= \sum_{i=1}^{c}\lambda_{ij}\sum_{k=1}^{n}(u_{ijk})^m(x_{kj}-y_{ij})^2 \\
J_j^4 &= \sum_{i=1}^{c}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](x_{kj}-y_{ij})^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \\
J_j^5 &= \sum_{i=1}^{c}\lambda_{ij}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](x_{kj}-y_{ij})^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \qquad (3.42)
\end{aligned}
$$

Em ambos os casos, a dispersão intra-grupo para C_i de acordo com a variável j onde $J^l=\sum_{i=1}^{c}J_i^l=\sum_{j=1}^{p}(\sum_{i=1}^{c}J_{ij}^l)$ é dada como a seguir:

$$
\begin{aligned}
J_{ij}^1 &= \sum_{k=1}^{n}(u_{ijk})^m(x_{kj}-y_{ij})^2 \\
J_{ij}^2 &= \sum_{k=1}^{n}\left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m(x_{kj}-y_{ij})^2 \\
J_{ij}^3 &= \lambda_{ij}\sum_{k=1}^{n}(u_{ijk})^m(x_{kj}-y_{ij})^2 \\
J_{ij}^4 &= \sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](x_{kj}-y_{ij})^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \\
J_{ij}^5 &= \lambda_{ij}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](x_{kj}-y_{ij})^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \qquad (3.43)
\end{aligned}
$$

A soma de quadrados inter-grupo B^l mede a dispersão dos protótipos do grupo e, conseqüentemente, a distinção de todos os grupos. Para isso, é medida a distância entre os protótipos e o centróide global $\mathbf{z}^l$. De acordo com a definição de agrupamento, o objetivo de métodos de agrupamento é maximizar esta medida, a fim de encontrar grupos de tal forma que os objetos em diferentes grupos são mais dissimilares do que os que estão no mesmo grupo. Esta medida é dada como:

$$
\begin{aligned}
B^1 &= \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk})^m(y_{ij}-z_j^1)^2 \\
B^2 &= \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}\left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m(y_{ij}-z_j^2)^2 \\
B^3 &= \sum_{i=1}^{c}\sum_{j=1}^{p}\lambda_{ij}\sum_{k=1}^{n}(u_{ijk})^m(y_{ij}-z_j^3)^2 \\
B^4 &= \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](y_{ij}-z_j^4)^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \\
B^5 &= \sum_{i=1}^{c}\sum_{j=1}^{p}\lambda_{ij}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](y_{ij}-z_j^5)^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \qquad (3.44)
\end{aligned}
$$

Em relação ao grupo C_i, a soma de quadrados inter-grupo é decomposta nas seguintes medidas:

$$
\begin{aligned}
B_i^1 &= \sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk})^m(y_{ij}-z_j^1)^2 \\
B_i^2 &= \sum_{j=1}^{p}\sum_{k=1}^{n}\left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m(y_{ij}-z_j^2)^2 \\
B_i^3 &= \sum_{j=1}^{p}\lambda_{ij}\sum_{k=1}^{n}(u_{ijk})^m(y_{ij}-z_j^3)^2 \\
B_i^4 &= \sum_{j=1}^{p}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](y_{ij}-z_j^4)^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \\
B_i^5 &= \sum_{j=1}^{p}\lambda_{ij}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](y_{ij}-z_j^5)^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \qquad (3.45)
\end{aligned}
$$

Similarmente, a dispersão dos protótipos dos grupos em relação à variável j pode ser calculado como:

$$
\begin{aligned}
B_j^1 &= \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ijk})^m(y_{ij}-z_j^1)^2 \\
B_j^2 &= \sum_{i=1}^{c}\sum_{k=1}^{n}\left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m(y_{ij}-z_j^2)^2 \\
B_j^3 &= \sum_{i=1}^{c}\lambda_{ij}\sum_{k=1}^{n}(u_{ijk})^m(y_{ij}-z_j^3)^2 \\
B_j^4 &= \sum_{i=1}^{c}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](y_{ij}-z_j^4)^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \\
B_j^5 &= \sum_{i=1}^{c}\lambda_{ij}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](y_{ij}-z_j^5)^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \qquad (3.46)
\end{aligned}
$$

Por fim, a dispersão dos protótipos em relação ao grupo C_i e variável j tal que $B^l = \sum_{i=1}^{c} B_i^l = \sum_{j=1}^{p}(\sum_{i=1}^{c} B_{ij}^l)$ pode ser calculado como:

$$
\begin{aligned}
B_{ij}^1 &= \sum_{k=1}^{n}(u_{ijk})^m(y_{ij}-z_j^1)^2 \\
B_{ij}^2 &= \sum_{k=1}^{n}\left(\frac{u_{ijk}}{\gamma_{ijk}}\right)^m(y_{ij}-z_j^2)^2 \\
B_{ij}^3 &= \lambda_{ij}\sum_{k=1}^{n}(u_{ijk})^m(y_{ij}-z_j^3)^2 \\
B_{ij}^4 &= \sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](y_{ij}-z_j^4)^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \\
B_{ij}^5 &= \lambda_{ij}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m+b(t_{ijk})^\eta\right](y_{ij}-z_j^5)^2+\pi_{ij}(1-t_{ijk})^\eta\right\} \qquad (3.47)
\end{aligned}
$$

Proposição 3.3.2. *As seguintes relações acontecem para todo* $i(i=1,\ldots,c)$, $j(j=1,\ldots,p)$ *e* $l=1,2,3,4,5$*:*

$$T^l = J^l + B^l,\ T_i^l = J_i^l + B_i^l,\ T_j^l = J_j^l + B_j^l \text{ e } T_{ij}^l = J_{ij}^l + B_{ij}^l. \qquad (3.48)$$

Prova 3.3.2. A prova é dada em A.11. □

3.3.2 Índices de Interpretação

Interpretar a qualidade de uma partição depois de ter aplicado um algoritmo de agrupamento difuso é um problema importante na Análise de Agrupamento. Nesta seção, serão apresentadas versões dos índices apresentados por Celeux et al. (1989) para algoritmos de agruamento difuso multivariados.

3.3.2.1 *Índices de Interpretação de Partição*

O índice de heterogeneidade global R^l $(l = 1,2,3,4,5)$ é medido pela relação entre a medida inter-grupo B^l e a dispersão global T^l. Ela é dada como:

$$R^l = \frac{B^l}{T^l} \tag{3.49}$$

Esse índice tem seus valores entre 0 e 1. É igual a 0 quando a partição difusa é um único grupo ou quando protótipos difusos são iguais ao centróide global. Considera-se que é igual a 1, quando todos os grupos difusos têm apenas um único objeto. Um valor de R^l próximo de 0 significa uma má representação dos elementos de um grupo difuso C_i pelo seu protótipo. Neste caso, o método de agregação não poderia encontrar grupos bem separados. Por outro lado, o valor de R^l próximo de 1 significa que a partição possui grupos mais homogêneos e uma melhor representação dos elementos de um grupo difuso C_i por seu protótipo.

Em relação a uma única variável, o índice de heterogeneidade global que mede a proporção de dispersão global e o índice inter-grupo em relação a variável j é definida como:

$$COR^l(j) = \frac{B^l_j}{T^l_j} \tag{3.50}$$

Esse índice tem também os seus valores entre 0 e 1, onde um valor de $COR^l(j)$ próximo de 0 significa que, de acordo com a variável j, protótipos são muito parecidos com o centróide global e denota uma partição de má qualidade sobre a variável j. Um valor de $COR^l(j)$ próximo de 1 denota uma melhor qualidade da partição relativa variável j.

A contribuição relativa da variável j à medida inter-grupo B^l é dada por:

$$CTR^l(j) = \frac{B^l_j}{B^l} \tag{3.51}$$

Um alto valor de $CTR^l(j)$ (próximo de 1) indica que a variável j fornece uma importante contribuição para a separação dos representantes dos grupos. Por outro lado, valores mais próximos de 0 significa que os grupos são próximos uns dos outros de acordo com a variável j.

3.3.2.2 *Índices de Interpretação de Grupos*

A análise da qualidade de um grupo é também um problema importante depois de ter aplicado um algoritmo de agrupamento difuso.

Em relação a medidas baseadas na soma do quadrados T^l, J^l e B^l e sua contribuição para o grupo C_i (T_i^l, J_i^l e B_i^l), é possível calcular a contribuição relativa, respectivamente, como:

$$\begin{aligned} T^l(i) &= \frac{T_i^l}{T^l} \\ J^l(i) &= \frac{J_i^l}{J^l} \\ B^l(i) &= \frac{B_i^l}{B^l} \end{aligned} \tag{3.52}$$

Eles tem valores entre 0 e 1. Alto valor de $T^l(i)$ indica que o grupo C_i tem um elevado contribuição para heterogeneidade global. Alto valor de $J^l(i)$ indica que o grupo C_i é relativamente heterogêneo em comparação com outros grupos. Valores próximos a 1 de $B^l(i)$ indica que o grupo C_i é relativamente distante do centro global em comparação com outros grupos.

Em relação a uma única variável, a heterogeneidade de um grupo pode ser diferente para diferentes outros. Por conseguinte, o índice de heterogeneidade global anterior podem ser decomposto na seguinte expressão:

$$COR^l(i,j) = \frac{B_{ij}^l}{T_j^l} \tag{3.53}$$

Um alto valor de $COR^l(i,j)$ indica que a variável j tem um comportamento relativamente homogêneo dentro do grupo C_i.

A contribuição relativa da variável j à heterogeneidade no grupo C_i é dada como:

$$CTR^l(i,j) = \frac{B_{ij}^l}{B_i^l} \tag{3.54}$$

Um alto valor de $CTR^l(i,j)$ indica que a variável j fornece uma importante contribuição para a medida entre-grupo em relação ao grupo C_i.

Por fim, a contribuição relativa da variável j e grupo C_i para a medida inter-grupo é calculado como:

$$CE^l(i,j) = \frac{B_{ij}^l}{B^l} \tag{3.55}$$

Valores próximos a 1 de $CE^l(i, j)$ significa que a variável j e o grupo C_i fornece uma importante contribuição para a medida inter-grupo. Em outras palavras, é possível verificar qual variável está contribuindo para a separação do grupo C_i em comparação com outros grupos da partição.

4

EXPERIMENTOS E RESULTADOS

Nossa maior fraqueza está em desistir. O caminho mais certo para vencer é tentar mais uma vez.

—THOMAS EDISON

4.1 Introdução

A fim de avaliar o desempenho dos métodos propostos, um estudo comparativo com seis métodos de agrupamento difuso é realizado. O primeiro método é o Gustafson-Kessel (aqui abreviado por GK) proposto por Gustafson e Kessel (1979) que pode identificar grupos de diferentes formatos. O segundo é o conhecido FCM proposto por Bezdek (1981), o qual é capaz de identificar grupos esféricos, uma vez que utiliza a distância Euclidiana. O terceiro método foi proposto por Keller e Klawonn (2000) (aqui abreviado FCM-C^1), onde há um parâmetro que fornece informações sobre a influência de variáveis individuais sobre os grupos detectados. O quarto é o *Possibilistic Fuzzy C-Means* (abreviado por PFCM) que une tanto as características da abordagem possibilística quanto difusa (Pal et al., 2005). O quinto método é uma versão do FCM usando distâncias adaptativas (aqui chamado FCM-C^2) proposto por De Carvalho et al. (2006a) que é capaz de identificar grupos de diferentes formatos e tamanhos. O sexto foi proposto por Nazari et al. (2013) o qual é uma versão do FCM com uma ponderação por variável (aqui abreviado por W-FCM).

A avaliação dos agrupamentos resultantes fornecidos por cada método baseia-se no cálculo de um índice de validade de agrupamento externo do tipo difuso proposto por Hullermeier et al. (2012) que é uma variação e generalização do índice de Rand (Rand, 1971) (aqui o índice difuso é abreviado por FR). O índice de Rand difuso tem os seus valores no intervalo $[0, 1]$, onde

o valor 1 indica uma perfeita concordância entre as partições *a priori* e a resultante do método de agrupamento, enquanto que valores perto de 0 correspondem às propriedades de recuperação insuficientes. FR, de acordo com partições difusas P e Q é, dada como:

$$FR(P,Q) = 1 - \frac{\sum_{k=1}^{n}\sum_{k'=1}^{n}|E_P(\mathbf{x}_k,\mathbf{x}_{k'}) - E_Q(\mathbf{x}_k,\mathbf{x}_{k'})|}{n(n-1)/2}, \tag{4.1}$$

onde $E_P(\mathbf{x}_k,\mathbf{x}_{k'}) = 1 - \|\boldsymbol{\delta}_k - \boldsymbol{\delta}_{k'}\|$ e $\boldsymbol{\delta}_k = (\delta_{1k},\ldots,\delta_{ik},\ldots,\delta_{ck})$ é um vetor de graus de pertinência por grupo do objeto $\mathbf{x}_k$ (similar definição $\boldsymbol{\delta}_{k'}$ com respeito ao objeto $\mathbf{x}_{k'}$ e $E_Q(\mathbf{x}_k,\mathbf{x}_{k'})$ para a partição Q). Aqui, a métrica $\|\cdot\|$ em $[0,1]^c$ que leva valores a $[0,1]$ é dada como $d(\boldsymbol{\delta}_k,\boldsymbol{\delta}_{k'}) = \frac{1}{c}\sum_{i=1}^{c}(\delta_{ik} - \delta_{ik'})^2$.

A motivação da utilização do índice difuso surge da necessidade de avaliar partições difusas de uma forma mais apropriada. O uso do índice de Rand original (para partições rígidas) na comparação de partições difusas poderia levar ao mesmo resultado, uma vez que é levado em conta somente a assocação 1 para a classe com o mais alto grau de pertinência e 0 para as demais. Como exemplo, considere a comparação entre os pares (P,Q_1) e (P,Q_2). A utilização do índice rígido produziria o mesmo resultado, ainda que a matriz Q_1 seja mais semelhante a P.

$$P = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix},$$

$$Q_1 = \begin{bmatrix} 0{,}99 & 0{,}94 & 0{,}90 & 0{,}08 & 0{,}91 & 0{,}01 & 0{,}02 & 0{,}00 \\ 0{,}01 & 0{,}06 & 0{,}10 & 0{,}92 & 0{,}09 & 0{,}99 & 0{,}98 & 1{,}00 \end{bmatrix},$$

$$Q_2 = \begin{bmatrix} 0{,}59 & 0{,}54 & 0{,}51 & 0{,}48 & 0{,}51 & 0{,}41 & 0{,}42 & 0{,}40 \\ 0{,}41 & 0{,}46 & 0{,}49 & 0{,}52 & 0{,}49 & 0{,}59 & 0{,}58 & 0{,}60 \end{bmatrix}.$$

Isto implica na perda de informação (Campello, 2007), descartando a utilização de índices rígidos para os métodos considerados neste trabalho. Diversos índices difusos foram propostos (Campello, 2007; Frigui et al., 2007; Brouwer, 2009; Anderson, 2010), entretanto, segundo Hullermeier et al. (2012), estes índices nem sempre atendem às propriedades tais como Reflexividade, Separação, Simetria e Desigualdade Triangular. Desta forma, o índice FR torna-se mais adequado à avaliação de partições difusas.

Conjuntos de dados sintéticos foram criados para avaliar o desempenho dos métodos em diferentes configurações. Além disso, conjuntos de dados do repositório de aprendizado de máquina da UCI (Bache e Lichman, 2013) foram utilizados para comparar os métodos. Selecionado um conjunto de dados da UCI, índices de partição de grupo foram calculados usando as partições resultantes dos métodos multivariados. O parâmetro m dos métodos difusos

utilizados neste trabalho foi fixado com o valor igual a 2. O número de grupos utilizado como parâmetro dos algoritmos é igual ao da partição original. Aqui, para calcular o índice FR segundo os resultados dos métodos multivariados, os graus de pertinência por grupo são calculados usando a função de agregação (Equação 3.7 e 3.13). Apêndice B apresenta os testes de hipótese relativos aos resultados apresentados para dados sintéticos e reais.

4.2 Dados Sintéticos

No contexto de conjunto de dados sintéticos, foram criados quatro cenários para avaliar o desempenho dos métodos. O primeiro cenário trata de configurações onde as partições são formadas por classes elípticas ou esféricas de tamanhos iguais ou diferentes. No segundo cenário, diferentemente do primeiro, as partições possuem tanto classes elípticas quanto esféricas, isto é, existe uma heterogeneidade de formatos. No terceiro, os métodos são avaliados em configurações onde as variáveis estão correlacionadas. Por fim, o quarto cenário contém configurações com classes esféricas e elípticas e diferentes percentuais de dados ruidosos a fim de avaliar o desempenho dos métodos que usam a abordagem possibilística.

Para os dados sintéticos, experimento Monte Carlo foi realizado com 100 réplicas. Este número é comumente utilizado na literatura de agrupamento (De Souza e De Carvalho, 2004; De Carvalho et al., 2006a,b,c; Ferreira e De Carvalho, 2014). Cada réplica é construída baseando-se nos parâmetros de cada classe que são arranjadas a partir de distribuições normais bi-variadas. Para cada réplica, um método de agrupamento foi aleatoriamente executado 50 vezes. O melhor resultado de acordo com a função objetivo foi selecionado e o índice FR foi calculado comparando a partição conhecida *a priori* com a obtida por cada método de agrupamento. Após a execução das 100 réplicas, a média e o desvio padrão do índice foram calculados.

4.2.1 Formato único das classes

Neste estudo, quatro conjuntos de dados no $\Re^2$ são considerados. Cada um contém três classes arranjadas de acordo com uma distribuição normal bivariada, cujos componentes são variáveis independentes. A principal diferença entre uma configuração para outra é o formato e o volume das classes. As variáveis não possuem correlação, uma vez que os métodos multivariados e os da literatura adotados neste estudo supõem *a priori* que as variáveis são independentes. Tabelas 4.1 e 4.2 mostram os parâmetros das classes para os conjuntos de dados de pontos de 1-4.

Parâmetro	Configuração 1			Configuração 2		
	Classe 1	Classe 2	Classe 3	Classe 1	Classe 2	Classe 3
μ_1	5	15	18	0	30	12
μ_2	0	5	14	0	0	25
σ_1^2	81	9	25	100	49	16
σ_2^2	9	100	36	100	49	16
Cardinalidade	200	100	50	200	100	50

Tabela 4.1: Parâmetros para os conjuntos de dados 1 e 2.

Parâmetro	Configuração 3			Configuração 4		
	Classe 1	Classe 2	Classe 3	Classe 1	Classe 2	Classe 3
μ_1	0	15	15	0	15	-15
μ_2	0	5	-5	0	0	0
σ_1^2	100	100	100	16	16	16
σ_2^2	4	4	4	16	16	16
Cardinalidade	100	100	100	150	100	50

Tabela 4.2: Parâmetros para os conjuntos de dados 3 e 4.

De acordo com os parâmetros, cada conjunto de dados contém três classes seguindo as seguintes descrições:

- Conjunto de dados 1: classes elípticas de diferentes tamanhos;
- Conjunto de dados 2: classes esféricas de diferentes tamanhos;
- Conjunto de dados 3: classes elípticas de tamanhos iguais;
- Conjunto de dados 4: classes esféricas de tamanhos iguais;

Figuras 4.1 e 4.2 mostram estes conjuntos de dados.

Tabela 4.3 mostra os valores da média e desvio padrão (em parênteses) do índice FR obtido por GK, FCM, FCM-C[1], PFCM, FCM-C[2], W-FCM, MFCM, MFCM-G, MFCM-D, MPFCM e MPFCM-D para os conjuntos de dados 1-4. Os maiores valores de média do índice FR para cada conjunto de dados estão em negrito.

Esses resultados mostram que o índice FR para o MFCM é maior do que GK, FCM, FCM-C[1], PFCM, FCM-C[2] e W-FCM para os conjuntos de dados 1 e 3 com classes elípticas. No entanto, para todos os conjuntos de dados, MFCM-G e MFCM-D obteveram melhor qualidade de agrupamento de que o MFCM. MFCM-G é superior aos outros para os conjunto de dados 1 (classes elípticas de tamanhos diferentes) e 3 (classes elípticas de tamanhos idênticos). Por outro lado, MFCM-D é superior para os conjuntos de dados 2 (classes esféricas de tamanhos diferentes)

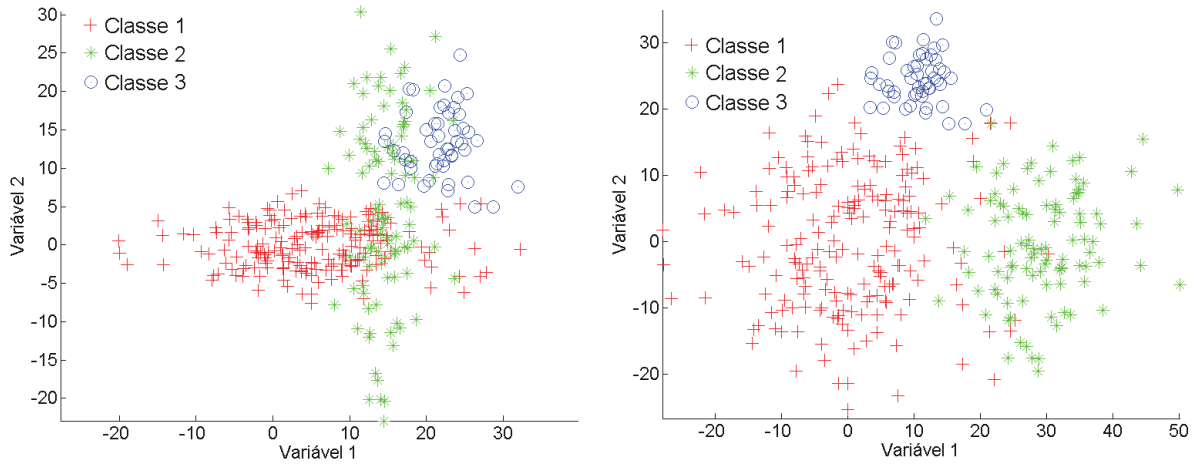

Figura 4.1: Conjunto de dados 1 (esquerda) e 2 (direita).

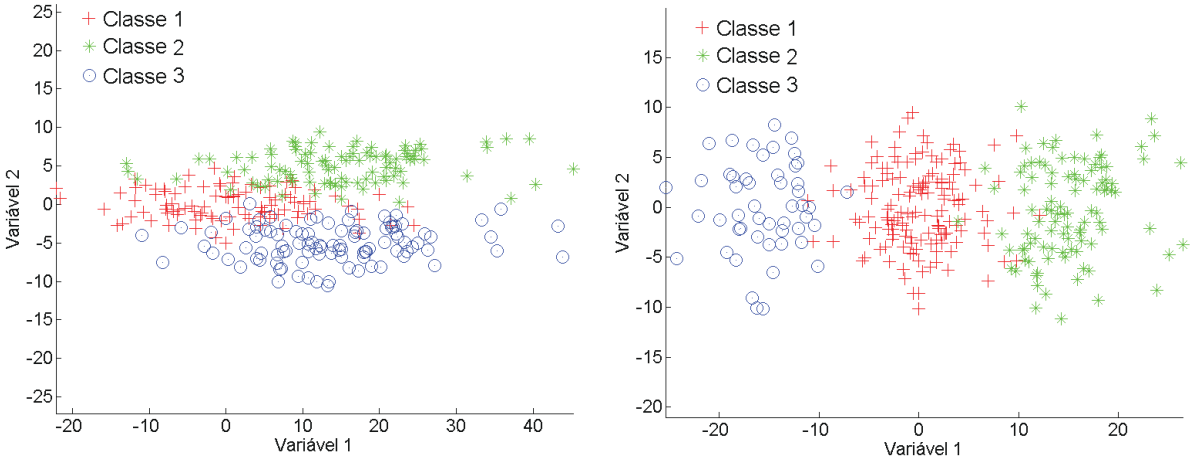

Figura 4.2: Conjunto de dados 3 (esquerda) e 4 (direita).

Conjunto	GK	FCM	FCM-C[1]	PFCM	FCM-C[2]	W-FCM	MFCM	MFCM-G	MFCM-D	MPFCM	MPFCM-D
1	0,5085 (0,0115)	0,4602 (0,0120)	0,4755 (0,0102)	0,4623 (0,0106)	0,5072 (0,0115)	0,4648 (0,0116)	0,5780 (0,0148)	**0,6883** (0,0183)	0,5840 (0,0147)	0,5405 (0,0167)	0,5615 (0,0094)
2	0,6184 (0,0111)	0,6351 (0,0107)	0,6224 (0,0113)	0,6366 (0,0120)	0,6172 (0,0117)	0,6283 (0,0107)	0,4856 (0,0104)	0,5126 (0,0248)	**0,6897** (0,0124)	0,4516 (0,0154)	0,5100 (0,0067)
3	0,6306 (0,0176)	0,4831 (0,0141)	0,5383 (0,0155)	0,4666 (0,0162)	0,6309 (0,0169)	0,5122 (0,0126)	0,6787 (0,0165)	**0,7956** (0,0149)	0,7232 (0,0134)	0,6642 (0,0175)	0,6625 (0,0110)
4	0,5474 (0,0181)	0,5628 (0,0158)	0,5566 (0,0153)	0,5652 (0,0131)	0,5467 (0,0180)	0,3409 (0,0180)	0,3647 (0,0097)	0,3916 (0,0300)	**0,6999** (0,0152)	0,2932 (0,0169)	0,4961 (0,0091)

Tabela 4.3: Média (maiores valores em negrito) e desvio padrão (em parênteses) do índice FR para os métodos de agrupamento e conjunto de dados 1-4.

e 4 (classes esféricas de tamanhos idênticos). Quanto aos métodos multivariados baseados na abordagem possibilística (MPFCM e MPFCM-D), eles foram superiores aos métodos presentes

na literatura para os conjuntos de dados 1 e 3. Além disso, MPFCM-D obteve melhores resultados do que sua versão não ponderada para todos os conjuntos, exceto no 3 onde eles foram estatisticamente iguais.

Em conclusão, o método MFCM mostrou ser adequado para configurações onde as classes têm variáveis de diferentes dispersões. Nota-se também que a utilização das versões ponderados de MFCM foi importante para o aumento da qualidade do agrupamento. O método MFCM-D é a melhor opção com grupos esféricos e MFCM-G com as elípticas. Além disso, MFCM é a pior opção entre todos os métodos onde as configurações das classes são esféricas, assim sua versão poderada com base nas dispersões é essencial para melhorar a qualidade de agrupamento. Os baseados na abodagem possibilística também obtiveram uma melhora qualidade de agrupamento quando as variáveis possuem diferentes dispersões e a versão poderada MPFCM-D mostrou ser tão bom ou melhor que sua versão não poderada. A Seção 4.2.2 apresenta uma avaliação dos métodos com uma mistura de grupos elípticos e esféricos presentes no conjunto de dados.

4.2.2 Mistura de Formas

Na Seção 4.2.1, o estudo considera que os conjuntos de dados com classes de mesmo formato, isto é, todas as classes são esféricas ou elípticas. Aqui, três conjuntos de dados com 4 classes de igual cardinalidade foram criados. O objetivo deste estudo é investigar o desempenho dos métodos com configurações contendo classes de diferentes formas. Os parâmetros destas classes são apresentados em Tabela 4.4 e 4.5.

Parâmetro	Configuração 5				Configuração 6			
	Classe 1	Classe 2	Classe 3	Classe 4	Classe 1	Classe 2	Classe 3	Classe 4
μ_1	5	15	10	3	5	15	12	7
μ_2	0	5	-7	15	0	5	-12	17
σ_1^2	81	9	49	25	81	9	16	25
σ_2^2	9	100	16	25	9	100	16	25
Cardinalidade	50	50	50	50	50	50	50	50

Tabela 4.4: Parâmetros para os conjuntos de dados 5 e 6.

Segundo os parâmetros apresentados, cada conjunto de dados contém quatro classes seguindo as seguintes descrições:

- Conjunto de dados 5: 3 classes elípticas e 1 esférica;
- Conjunto de dados 6: 2 classes elípticas e 2 esféricas;

Parâmetro	Configuração 7			
	Classe 1	Classe 2	Classe 3	Classe 4
μ_1	0	18	-18	0
μ_2	0	0	0	-12
σ_1^2	12	20	16	81
σ_2^2	12	20	16	20
Cardinalidade	50	50	50	50

Tabela 4.5: Parâmetros para o conjunto de dados 7.

- Conjunto de dados 7: 1 classe elíptica e 3 esféricas;

Figuras 4.3 e 4.4 mostram estes conjuntos de dados.

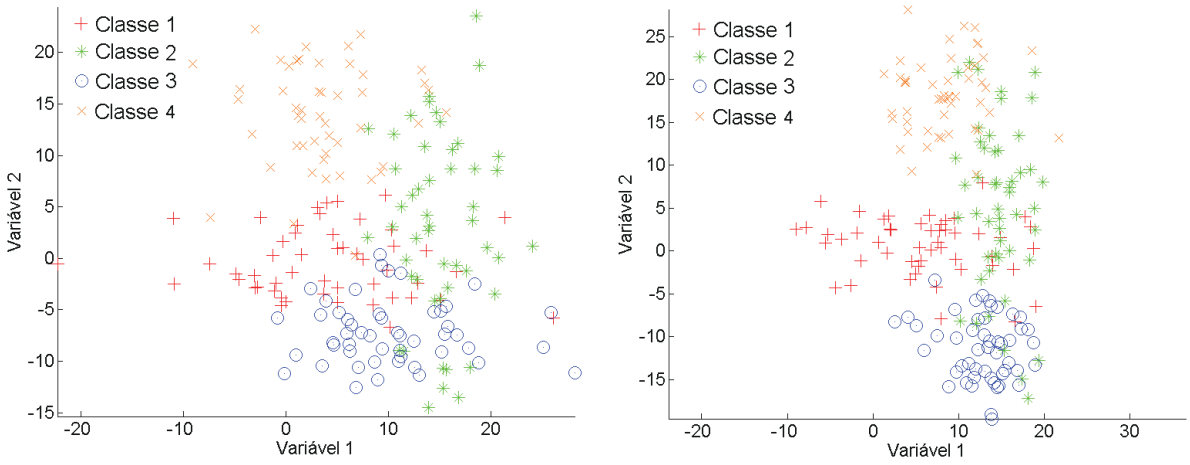

Figura 4.3: Conjunto de dados 5 (esquerda) e 6 (direita).

A partir da Tabela 4.6, podem-se extrair algumas observações:

1. Os métodos GK, FCM, FCM-C[1], FCM-C[2], W-FCM, MFCM-D e MPFCM-D melhoraram o seu desempenho quando o número de classes esféricas presentes no conjunto de dados aumenta. Por outro lado, o MFCM tem o pior desempenho quando existem mais classes esféricas do que elípticas (conjunto de dados 7).

2. O índice FR para MFCM-G degrada quando o número de classes esféricas presentes no conjunto de dados aumenta. No entanto, este método é superior aos outros considerados neste estudo para os conjuntos de dados 5 e 6.

3. Os métodos MPFCM e MPFCM-D foram superiores a sua versão não multivariada PFCM. Além disso, MPFCM-D obteve melhores resultados que o MPFCM para os conjuntos conjuntos de dados 5, 6 e 7.

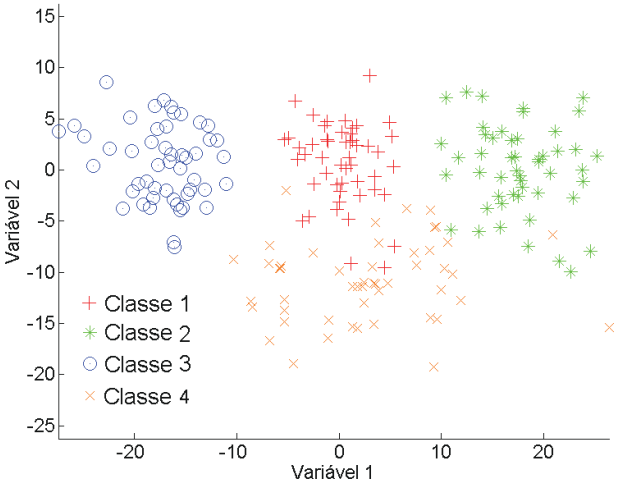

Figura 4.4: Conjunto de dados 7.

4. O uso de pesos é relevante para obter uma melhor qualidade de agrupamento uma vez que estes métodos superaram os métodos sem ponderação para as variáveis para todos os conjuntos de dados. MFCM-D deve ser preferido quando existir mais classes esféricas do que as elípticas.

Conjunto	GK	FCM	FCM-C[1]	PFCM	FCM-C[2]	W-FCM	MFCM	MFCM-G	MFCM-D	MPFCM	MPFCM-D
5	0,5317 (0,0145)	0,4839 (0,0181)	0,5105 (0,0104)	0,4843 (0,0093)	0,4973 (0,0196)	0,5034 (0,0082)	0,5576 (0,0092)	**0,6966** (0,0186)	0,5871 (0,0192)	0,5158 (0.0055)	0,5263 (0,0050)
6	0,5772 (0,0052)	0,5413 (0,0241)	0,5807 (0,0127)	0,4463 (0.0362)	0,5432 (0,0196)	0,5240 (0,0096)	0,5798 (0,0124)	**0,6766** (0,0206)	0,6502 (0,0216)	0,5221 (0,0195)	0,5438 (0,0152)
7	0,6394 (0,0060)	0,6056 (0,0366)	0,6128 (0,0140)	0,4843 (0,0264)	0,6008 (0,0536)	0,6029 0,0058	0,5424 (0,0213)	0,6252 (0,0448)	**0,6732** (0,0145)	0,5274 (0,0270)	0,5767 (0,0166)

Tabela 4.6: Média e desvio padrão (em parênteses) do índice FR para os métodos de agrupamento e conjuntos de dados 5-7.

4.2.3 Variáveis Correlacionadas

Nas seções 4.2.1 e 4.2.2, as partições foram construídas de forma que as variáveis fossem independentes. Nesta seção, o objetivo deste cenário de conjuntos de dados é avaliar o desempenho dos métodos quando existe dependência entre as variáveis. Para isto, foram criados quatro conjuntos com variáveis correlacionadas. Os conjuntos de dados possuem classes arranjadas de acordo com uma distribuição normal bivariada. O primeiro conjunto de dados é semelhante à Configuração 1 onde as classes são elípticas de tamanhos diferentes de forma que a principal diferença está na correlação entre as variáveis. O segundo e terceiro conjuntos

são similares entre si, exceto pelo fato de uma ser desbalanceada e outra não, respectivamente. O quarto conjunto possui duas classes e, diferentemente dos outros conjuntos, utilizou-se uma matriz de rotação cujos ângulos aplicados foram 0º, 5º, 10º, 15º, 20º, 25º, 30º, 35º, 40º e 45º. Os parâmetros destas classes são apresentados nas tabelas 4.7 e 4.8.

Parâmetro	Configuração 8			Configuração 9		
	Classe 1	Classe 2	Classe 3	Classe 1	Classe 2	Classe 3
μ_1	5	15	21	0	15	30
μ_2	0	5	14	0	0	0
σ_1^2	81	9	16	81	81	81
σ_2^2	9	100	25	81	81	81
ρ	$0,9$	$0,9$	$-0,8$	$0,9$	$0,9$	$0,9$
Cardinalidade	200	100	50	200	100	50

Tabela 4.7: Parâmetros para os conjuntos com dados 8 e 9.

Parâmetro	Configuração 10			Configuração 11	
	Classe 1	Classe 2	Classe 3	Classe 1	Classe 2
μ_1	0	15	30	$0,5$	$1,5$
μ_2	0	0	0	$1,5$	$1,5$
σ_1^2	81	81	81	$0,0277$	$0,0277$
σ_2^2	81	81	81	$0,25$	$0,25$
ρ	$0,9$	$0,9$	$0,9$	–	–
Cardinalidade	100	100	100	100	100

Tabela 4.8: Parâmetros para os conjuntos com dados 10 e 11.

Figuras 4.5 e 4.6 apresentam as configurações com variáveis correlacionadas.

Tabelas 4.9 e 4.10 mostram os valores da média e desvio padrão (em parênteses) do índice FR obtido pelos métodos GK, FCM, FCM-C[1], PFCM, FCM-C[2], W-FCM, MFCM, MFCM-G, MFCM-D, MPFCM e MPFCM-D para os conjuntos de dados 8-11. Os maiores valores de média do índice FR para cada conjunto de dados estão em negrito.

Conjunto	GK	FCM	FCM-C[1]	PFCM	FCM-C[2]	W-FCM	MFCM	MFCM-G	MFCM-D	MPFCM	MPFCM-D
8	$0,4924$ ($0,0145$)	$0,4614$ ($0,0117$)	$0,4831$ ($0,0131$)	$0,4633$ ($0,0121$)	$0,5049$ ($0,0145$)	$0,4620$ ($0,0126$)	$0,5364$ ($0,0160$)	**0,6540** ($0,0185$)	$0,5608$ ($0,0167$)	$0,5334$ ($0,0118$)	$0,5620$ ($0,0134$)
9	$0,4317$ ($0,0157$)	$0,4602$ ($0,0120$)	$0,4512$ ($0,0132$)	$0,4617$ ($0,0128$)	$0,4466$ ($0,0166$)	$0,3967$ ($0,0095$)	$0,3804$ ($0,0128$)	$0,3828$ ($0,0284$)	**0,5332** ($0,0141$)	$0,3099$ ($0,0103$)	$0,3942$ ($0,0246$)
10	$0,4075$ ($0,0115$)	$0,4072$ ($0,0119$)	$0,4107$ ($0,0118$)	$0,4087$ ($0,0175$)	$0,4105$ ($0,0142$)	$0,3322$ ($0,0154$)	$0,3542$ ($0,0165$)	$0,3515$ ($0,0513$)	**0,5046** ($0,0133$)	$0,2317$ ($0,0155$)	$0,3307$ ($0,0126$)

Tabela 4.9: Média e desvio padrão (entre parênteses) do índice FR (maiores valores em negrito) para conjunto de dados 8, 9 e 10.

Segundo os resultados apresentados na Tabela 4.9, o método MFCM-G continua sendo superior aos demais no conjunto de dados 8. As versões ponderadas MFCM-G e MFCM-D

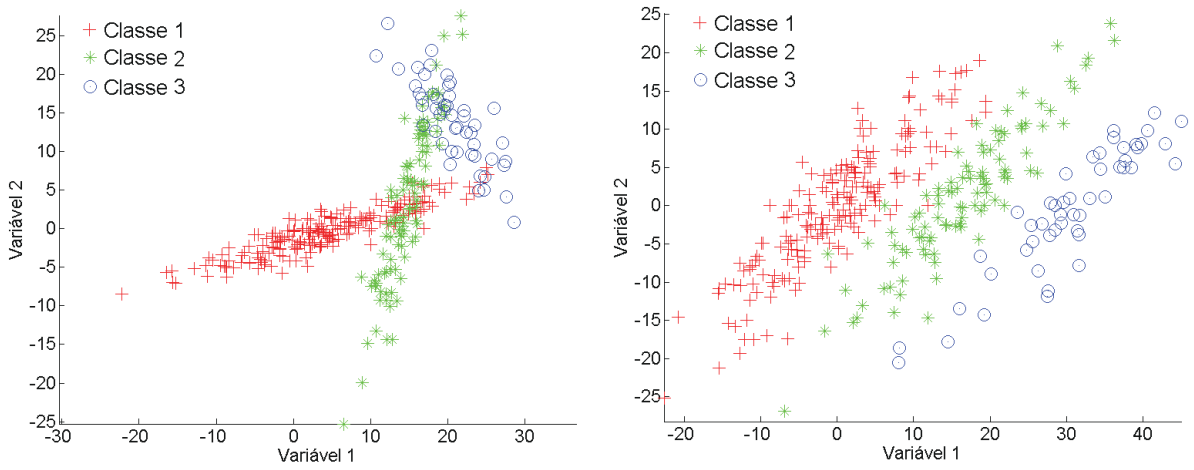

Figura 4.5: Conjunto de dados 8 (esquerda) e 9 (direita).

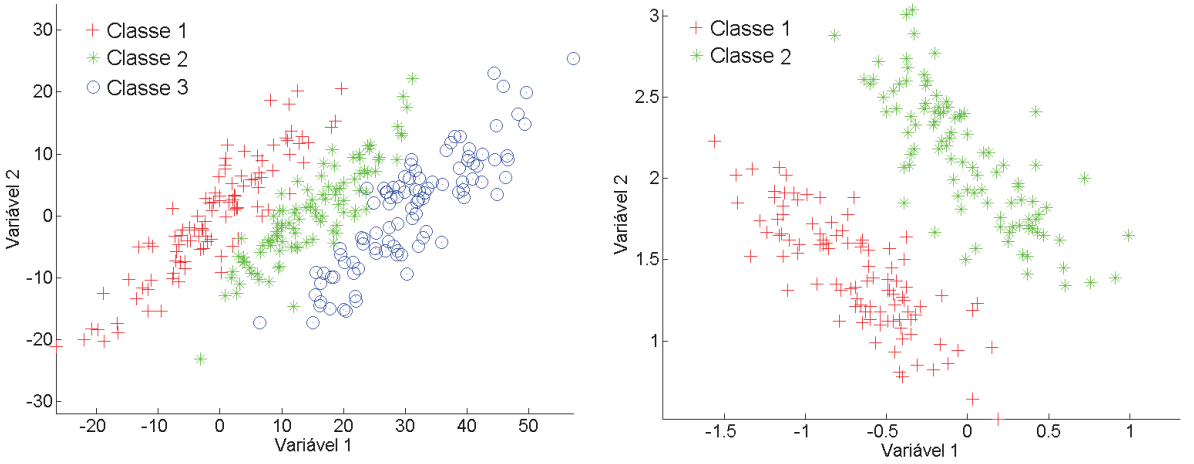

Figura 4.6: Conjunto de dados 10 (esquerda) e 11 (direita).

obtiveram melhores resultados que o método MFCM. Similarmente, o método MPFCM-D foi superior ao MPFCM. No conjunto de dados 9, o método MFCM-D foi a melhor opção. Pode-se notar também que os métodos multivariados mostraram ser sensíveis à correlação. Neste conjunto de dados, MFCM-G foi tão bom quanto MFCM. Já no conjunto de dados 10, nota-se que o desempenho dos métodos foi reduzido em comparação ao conjunto 9 onde as classes estão balanceadas. Neste conjunto, MFCM-D foi a melhor opção e MPFCM-D foi superior a sua versão não ponderada MPFCM.

Quanto aos resultados apresentados na Tabela 4.10, pode-se notar que as versões ponderadas MFCM-G e MFCM-D foram superiores a sua versão não ponderada MFCM. O método multivariado com ponderação na distância MFCM-D obteve melhor qualidade de agrupamento

Ângulo	GK	FCM	FCM-C[1]	PFCM	FCM-C[2]	W-FCM	MFCM	MFCM-G	MFCM-D	MPFCM	MPFCM-D
0º	0,7834 (0,0194)	0,5121 (0,0210)	0,6467 (0,0202)	0,5108 (0,0210)	0,7835 (0,0195)	0,3713 (0,0128)	0,4889 (0,0306)	0,6132 (0,0398)	**0,8481** (0,0196)	0,4526 (0,0324)	0,5558 (0,0326)
5º	0,7750 (0,0211)	0,5121 (0,0210)	0,6399 (0,0195)	0,5108 (0,0210)	0,7750 (0,0211)	0,3711 (0,0151)	0,4994 (0,0345)	0,5991 (0,0588)	**0,8422** (0,0145)	0,4361 (0,0397)	0,5670 (0,0223)
10º	0,7468 (0,0221)	0,5121 (0,0210)	0,6282 (0,0177)	0,5108 (0,0210)	0,7454 (0,0196)	0,3772 (0,0145)	0,5012 (0,0376)	0,5833 (0,0407)	**0,8172** (0,0192)	0,4082 (0,0386)	0,5464 (0,0096)
15º	0,7070 (0,0228)	0,5121 (0,0210)	0,6085 (0,0178)	0,5108 (0,0210)	0,7035 (0,0243)	0,3847 (0,0194)	0,4940 (0,0497)	0,5343 (0,0360)	**0,7894** (0,0207)	0,3899 (0,0315)	0,4429 (0,0141)
20º	0,6536 (0,0232)	0,5121 (0,0210)	0,5855 (0,0199)	0,5108 (0,0210)	0,6551 (0,0211)	0,4076 (0,0291)	0,4814 (0,0561)	0,4772 (0,0418)	**0,7399** (0,0273)	0,3639 (0,0359)	0,4276 (0,0200)
25º	0,6098 (0,0234)	0,5121 (0,0210)	0,5593 (0,0209)	0,5108 (0,0210)	0,6087 (0,0240)	0,4451 (0,0348)	0,4545 (0,0678)	0,4374 (0,0506)	**0,6835** (0,0264)	0,3400 (0,0277)	0,3832 (0,0543)
30º	0,5751 (0,0224)	0,5121 (0,0210)	0,5422 (0,0220)	0,5108 (0,0210)	0,5738 (0,0222)	0,4781 (0,0301)	0,4180 (0,0718)	0,4169 (0,0476)	**0,6103** (0,0325)	0,3230 (0,0254)	0,3520 (0,0052)
35º	**0,5419** (0,0193)	0,5121 (0,0210)	0,5279 (0,0229)	0,5108 (0,0210)	0,5399 (0,0238)	0,4978 (0,0283)	0,4299 (0,0792)	0,4021 (0,0465)	0,2624 (0,1154)	0,3081 (0,0259)	0,3276 (0,0120)
40º	**0,5167** (0,0212)	0,5121 (0,0210)	0,5162 (0,0204)	0,5108 (0,0210)	0,5148 (0,0246)	0,5033 (0,0243)	0,4443 (0,0838)	0,3930 (0,0481)	0,2471 (0,0488)	0,3009 (0,0278)	0,3042 (0,0195)
45º	**0,5134** (0,0222)	0,5121 (0,0210)	0,5124 (0,0205)	0,5108 (0,0210)	0,5123 (0,0215)	0,5083 (0,0225)	0,4505 (0,0808)	0,3899 (0,0434)	0,3541 (0,0351)	0,3008 (0,0255)	0,3069 (0,0254)

Tabela 4.10: Média e desvio padrão (entre parênteses) do índice FR (maiores valores em negrito) para conjunto de dados 11.

em relação aos demais métodos quando o ângulo utilizado foi até 30º. Por outro lado, quando o ângulo ficou acima de 30º, este método teve uma queda no seu desempenho. Além disso, comparando o seu desempenho ao longo da variação do ângulo, nota-se que MFCM-D é o mais sensível dentre os métodos multivariados. Em contraste, o método MFCM mostrou ser o menos sensível à variação do ângulo dentre todos os métodos multivariados. Em relação aos métodos baseados na abordagem possibilística, o método multivariado MPFCM-D foi superior ao tradicional PFCM para baixos valores de ângulo (0º, 5º e 10º). Além disso, MPFCM-D mostrou ser superior a sua versão não ponderada MPFCM para valores de ângulo até 35º e para os valores 40º e 45º obtiveram resultados similares. Isso mostra que a ponderação torna os métodos MFCM e MPFCM menos sensíveis a variáveis correlacionadas, o que é compreensível já que os pesos supõem que não existe dependência entre elas.

4.2.4 Dados Ruidosos

Os conjuntos sintéticos com dados ruidosos apresentados nessa seção foram construídos baseando-se em configurações utilizadas em artigos presentes na literatura. Hu et al. (2010) acrescentou dados ruidosos em dados reais de acordo com os percentuais: 6%, 12%, 18%, 24% e 30%. Xiong (2006) alega que o percentual de ruído pode variar de acordo com o domínio ao qual o conjunto de dados pertence. Por exemplo, no comércio, os dados ruidosos variam em torno de 5%. Mas em outros casos essa quantidade pode ser relativamente grande, como em dados de proteínas onde pode passar de 30%. Leski (2016) criou configurações onde os ruídos estão espalhados aleatoriamente de forma que sigam distribuições uniformes; em suas

configurações, os dados aberrantes estão no intervalo: $[-300, 50]$ e $[-200, 1000]$. Ji et al. (2014) usa um conjunto de dados simulados com 37,5% de ruído. Atla (2011) fez experimentos onde a proporção de dados ruidosos varia de acordo com os seguintes percentuais: 0%, 10%, 20%, 30%, 40% e 50%.

Neste trabalho, os conjuntos com dados ruidosos possuem três classes arranjadas de acordo com uma distribuição normal bivariada, cujos componentes são variáveis independentes. No primeiro conjunto as classes são esféricas, enquanto no segundo conjunto as classes são elípticas. Em ambos os conjuntos, foram acrescentados ruídos arranjados de acordo com uma distribuição uniforme bivariada variando no intervalo $[-100, 50]$ para ambas as variáveis e configurações. Foram feitos experimentos com os seguintes percentuais de dados ruidosos: 0%, 5%, 10%, 15%, 20%, 25% e 30%. Os parâmetros destas classes são apresentados na Tabela 4.11.

Parâmetro	Configuração 12			Configuração 13		
	Classe 1	Classe 2	Classe 3	Classe 1	Classe 2	Classe 3
μ_1	-16	-8	0	7	0	12
μ_2	-5	8	16	20	-6	0
σ_1^2	20	13	6	2	50	50
σ_2^2	20	13	6	50	2	2
Cardinalidade	50	100	200	100	100	100

Tabela 4.11: Parâmetros para os conjuntos com dados 12 e 13.

Figuras 4.7 e 4.8 apresentam as configurações com dados ruidosos segundo o percentual 0% e 30%.

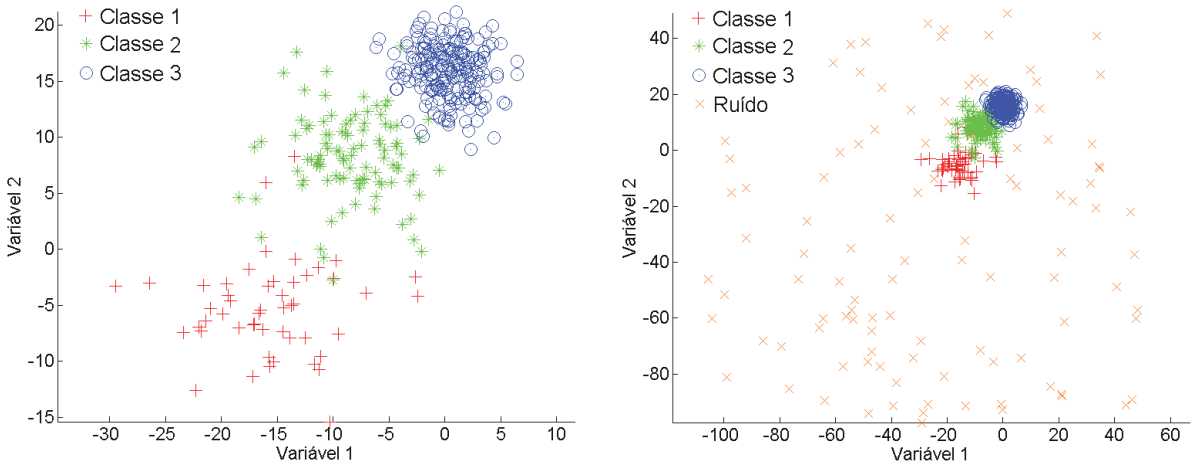

Figura 4.7: Conjunto de dados 12 com 0% de ruido (esquerda) e 30% (direita).

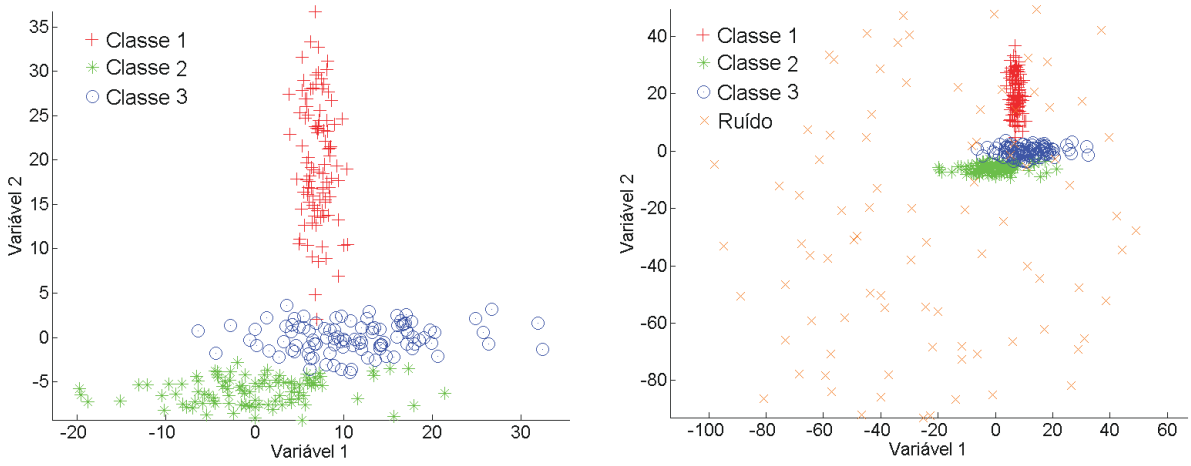

Figura 4.8: Conjunto de dados 13 com 0% de ruido (esquerda) e 30% (direita).

Tabelas 4.12 e 4.13 mostram os valores da média e desvio padrão (em parênteses) do índice FR obtido pelos métodos GK, FCM, FCM-C[1], PFCM, FCM-C[2], W-FCM, MFCM, MFCM-G, MFCM-D, MPFCM e MPFCM-D para os conjuntos de dados 12 e 13. Os maiores valores de média do índice FR para cada conjunto de dados estão em negrito.

Percentual	GK	FCM	FCM-C[1]	PFCM	FCM-C[2]	W-FCM	MFCM	MFCM-G	MFCM-D	MPFCM	MPFCM-D
0%	0,7057 (0,0139)	0,7056 (0,0140)	0,7056 (0,0142)	0,6810 (0,0153)	0,7056 (0,0139)	0,6994 (0,0134)	0,6692 (0,0238)	0,7416 (0,0263)	**0,8520** (0,0200)	0,7208 (0,0190)	0,7218 (0,0193)
5%	0,6878 (0,0166)	0,6900 (0,0152)	0,6912 (0,0152)	0,6844 (0,0159)	0,6882 (0,0162)	0,6864 (0,0140)	0,6539 (0,0212)	0,7376 (0,0283)	**0,8008** (0,0284)	0,7281 (0,0181)	0,7288 (0,0185)
10%	0,6846 (0,0151)	0,6902 (0,0157)	0,6878 (0,0146)	0,6839 (0,0151)	0,6862 (0,0159)	0,6824 (0,0142)	0,6442 (0,0403)	0,7125 (0,0300)	**0,7666** (0,0367)	0,7242 (0,0176)	0,7245 (0,0175)
15%	0,6832 (0,0163)	0,6872 (0,0160)	0,6863 (0,0158)	0,6841 (0,0151)	0,6832 (0,0153)	0,6829 (0,0142)	0,6291 (0,0328)	0,6758 (0,0307)	**0,7551** (0,0442)	0,7247 (0,0165)	0,7257 (0,0169)
20%	0,6330 (0,1025)	0,5875 (0,1288)	0,6104 (0,1200)	0,6831 (0,0148)	0,6124 (0,1163)	0,6502 (0,0876)	0,6227 (0,0238)	0,6668 (0,0636)	0,7241 (0,0387)	**0,7305** (0,0213)	**0,7320** (0,0212)
25%	0,4308 (0,0107)	0,4255 (0,0276)	0,4325 (0,0517)	0,6782 (0,0148)	0,4308 (0,0109)	0,4344 (0,0618)	0,6099 (0,0261)	0,6652 (0,0677)	0,6566 (0,0462)	**0,7242** (0,0201)	**0,7248** (0,0197)
30%	0,4297 (0,0102)	0,4227 (0,0094)	0,4228 (0,0102)	0,6748 (0,0144)	0,4295 (0,0098)	0,4185 (0,0105)	0,5831 (0,0248)	0,6663 (0,0723)	0,6003 (0,0741)	**0,7263** (0,0208)	**0,7276** (0,0201)

Tabela 4.12: Média e desvio padrão (entre parênteses) do índice FR (maiores valores em negrito) para conjunto com dados ruidosos e classes esféricas.

De acordo com os resultados apresentados na Tabela 4.12, pode-se notar que as versões poderadas MFCM-G e MFCM-D mostraram ser superiores que sua versão não poderada MFCM para todos os percentuais. Em especial, para valores de 0% a 15% de dados ruidosos, o método MFCM-D obteve melhores resultados quando comparado com outros métodos. Os métodos baseados na abordagem possibilística (PFCM, MPFCM e MPFCM-D) mostraram ser estáveis à variação do percentual de dados ruidosos. Além disso, os multivariados MPFCM e MPFCM-D conseguiram obter melhores resultados em relação aos outros métodos com o percentual a

partir de 20%. Quanto aos métodos multvariados MPFCM e MPFCM-D, apresentaram um comportamento similar para este conjunto de dados com classes esféricas.

Percentual	GK	FCM	FCM-C[1]	PFCM	FCM-C[2]	W-FCM	MFCM	MFCM-G	MFCM-D	MPFCM	MPFCM-D
0%	0,7670 (0,0660)	0,6073 (0,0180)	0,5693 (0,0316)	0,5774 (0,0203)	0,7733 (0,0136)	0,5723 (0,0208)	0,7284 (0,0279)	**0,8129** (0,0344)	0,7416 (0,0236)	0,6650 (0,0380)	0,6875 (0,0354)
5%	0,6598 (0,0396)	0,5897 (0,0190)	0,5876 (0,0361)	0,5816 (0,0212)	0,6616 (0,0401)	0,5703 (0,0207)	0,7205 (0,0580)	**0,7150** (0,0611)	0,5578 (0,0498)	0,6787 (0,0348)	0,6991 (0,0336)
10%	0,6035 (0,0660)	0,5732 (0,0391)	0,5326 (0,0553)	0,5803 (0,0196)	0,5972 (0,0688)	0,5635 (0,0286)	0,6551 (0,0436)	0,6311 (0,0809)	0,5413 (0,0398)	0,6756 (0,0329)	**0,6894** (0,0300)
15%	0,5117 (0,0200)	0,4846 (0,0263)	0,4908 (0,0282)	0,5734 (0,0193)	0,5111 (0,0201)	0,4881 (0,0309)	0,6178 (0,1249)	0,5737 (0,0851)	0,5104 (0,0637)	0,6728 (0,0327)	**0,6884** (0,0320)
20%	0,5102 (0,0143)	0,4802 (0,0106)	0,4957 (0,0220)	0,5744 (0,0224)	0,5102 (0,0157)	0,4803 (0,0143)	0,5082 (0,0707)	0,5393 (0,0570)	0,4996 (0,0658)	0,6696 (0,0376)	**0,6856** (0,0320)
25%	0,5168 (0,0175)	0,4750 (0,0115)	0,4907 (0,0187)	0,5678 (0,0201)	0,5151 (0,0188)	0,4757 (0,0151)	0,4859 (0,0148)	0,5376 (0,0312)	0,4911 (0,0663)	0,6662 (0,0341)	**0,6844** (0,0356)
30%	0,5166 (0,0194)	0,4750 (0,0124)	0,4920 (0,0203)	0,5584 (0,0190)	0,5160 (0,0208)	0,4753 (0,0148)	0,4840 (0,0255)	0,5420 (0,0328)	0,4627 (0,0760)	0,6624 (0,0332)	**0,6756** (0,0345)

Tabela 4.13: Média e desvio padrão (entre parênteses) do índice FR (maiores valores em negrito) para conjunto com dados ruidosos e classes elípticas.

Quanto à Tabela 4.13, para mais baixos percentuais de dados ruidosos (0% e 5%), o método multivariado com ponderação no grau de pertinência MFCM-G foi superior aos demais. De um modo mais geral, as versões ponderadas MFCM-G e MFCM-D obtiveram melhores resultados que sua versão não ponderada MFCM. Em relação aos métodos baseados na abordagem possibilística (PFCM, MPFCM e MPFCM-D), semelhantemente à configuração com classes esféricas, mostraram ser estáveis à variação do percentual de dados ruidosos. Os métodos multivariados MPFCM e MPFCM-D apresentaram melhores resultados aos demais quando o percentual foi superior a 5%. E para este conjunto de dados com classes elípticas, o método MPFCM-D obteve uma melhor qualidade de agrupamento que o MPFCM para todos os percentuais de dados ruidosos. Em relação aos resultados obtidos pelos métodos de agrupamento em ambos os conjuntos de dados, pode-se dizer que os métodos multivariados baseados na abordagem possibilística MPFCM e MPFCM-D foram superiores a sua versão não multivariada PFCM e mostraram ser menos sensíveis à variação do percentual de dados ruidosos do que os métodos que não usam a abordagem possibilística.

4.3 Conjunto de Dados Reais da UCI

Métodos de agrupamento difuso foram aplicadas para 24 conjuntos de dados reais da UCI (Bache e Lichman, 2013). A Tabela 4.14 apresenta o nome desses conjuntos assim como sua descrição: número de classes, número de objetos, a cardinalidade de objetos em cada classe e o número de variáveis. Esses conjuntos foram adotados nesse trabalho devido ao número relativamente baixo de objetos e variáveis.

Cada conjunto de dados reais foi padronizado de forma que os valores de cada variável estejam no intervalo $[0,1]$. Usando a técnica *Booststrap*, foram realizadas 1000 reamostragens de cada conjunto de dados reais. Para cada reamostragem, um método é executado, até a sua convergência, 200 vezes com inicializações aleatórias e o melhor resultado segundo a função objetivo é selecionado. Uma vez feita a seleção, o índice FR é calculado para cada método comparando a partição obtida pelo método e a partição conhecida *a priori*. Ao final da execução das 1000 reamostragens, a média e o desvio padrão são calculados. O parâmetro m dos métodos difusos foi fixado com o valor igual a 2.

Conjunto	# Classes	# Obj.	Cardinalidade	# Var.
Abalone	3	4177	1307, 1342, 1528	8
Banknote	2	1372	762, 610	4
Blood	2	748	570, 178	4
Breast	2	569	357, 212	30
BreastTissue	6	106	21, 15, 18, 16, 14, 22	9
Climate	2	540	46, 494	19
Ecoli	8	336	143, 77, 52, 35, 20, 5, 2, 2	7
Glass	6	214	70, 76, 17, 13, 9, 29	9
Haberman	2	306	225, 81	3
ImageSeg	7	210	30, 30, 30, 30, 30, 30, 30	19
Ionosphere	2	351	126, 225	34
Iris	3	150	50, 50, 50	4
Parkinsons	2	197	48, 149	22
Pima	2	768	500, 268	8
QSAR	2	1055	284, 771	41
Seeds	3	210	70, 70, 70	7
Sonar	2	208	98, 110	60
StatlogV	4	946	240, 240, 240, 246	18
Thyroid	3	215	150, 35, 30	5
Vertebral	2	310	210, 100	6
WDBC	2	569	212, 357	30
Wine	3	178	59, 71, 48	13
WineQlt	6	4898	10, 53, 681, 639, 200, 19	11
Yeast	10	1484	463, 5, 35, 44, 51, 163 244, 429, 20, 30	8

Tabela 4.14: Descrição de conjuntos de dados reais da UCI.

Tabela 4.15 mostra os valores da média e desvio padrão (em parênteses) do índice FR obtido pelos métodos GK, FCM, FCM-C[1], PFCM, FCM-C[2], W-FCM, MFCM, MFCM-G, MFCM-D, MPFCM e MPFCM-D para os conjuntos de dados reais da UCI. Os maiores valores de média do índice FR para cada conjunto de dados estão em negrito.

A partir da Tabela 4.15, nota-se que, dos 24 conjuntos de dados reais, em 17 deles os métodos multivariados obtiveram uma melhor qualidade de agrupamento. Desses 17 conjuntos,

Conjunto	GK	FCM	FCM-C[1]	PFCM	FCM-C[2]	W-FCM	MFCM	MFCM-G	MFCM-D	MPFCM	MPFCM-D
Abalone	0,3851 (0,0516)	0,3492 (0,0262)	0,3609 (0,0337)	0,2817 (0,0209)	0,3887 (0,0342)	0,3431 (0,0293)	**0,5045** (0,0456)	0,2140 (0,0558)	0,4902 (0,0254)	0,3869 (0,0227)	0,4095 (0,0342)
Banknote	0,0201 (0,0284)	0,0466 (0,0132)	0,0201 (0,0342)	0,0415 (0,0214)	0,0348 (0,0211)	0,0445 (0,0124)	**0,1341** (0,0128)	0,0441 (0,0252)	0,0089 (0,0314)	0,0145 (0,0213)	0,0362 (0,0241)
Blood	**0,4004** (0,0665)	0,3130 (0,0623)	0,3090 (0,0515)	0,3201 (0,0586)	0,2940 (0,0370)	0,3520 (0,0621)	0,0400 (0,0353)	0,2790 (0,0666)	0,1749 (0,0442)	0,2262 (0,0494)	0,2438 (0,0581)
Breast	0,4533 (0,0402)	0,3884 (0,0445)	0,4208 (0,0466)	0,3696 (0,0143)	0,4738 (0,0502)	0,3370 (0,0607)	0,6313 (0,2222)	0,3177 (0,1054)	**0,7661** (0,2106)	0,3718 (0,1641)	0,3956 (0,1579)
BreastTissue	0,6808 (0,0784)	0,6984 (0,0590)	0,6657 (0,0585)	0,6481 (0,0626)	0,6808 (0,0784)	0,6594 (0,0579)	0,7216 (0,1347)	0,7201 (0,0861)	**0,9910** (0,0160)	0,7307 (0,0316)	0,7380 (0,0299)
Climate	0,6467 (0,0213)	0,8085 (0,0231)	0,6467 (0,0124)	0,8070 (0,0184)	0,5609 (0,0241)	0,8002 (0,0245)	**0,9643** (0,0214)	0,9434 (0,0314)	0,9107 (0,0311)	0,9301 (0,0142)	0,9586 (0,0241)
Ecoli	0,7318 (0,0413)	0,7433 (0,0134)	0,6380 (0,0313)	**0,7338** (0,0134)	0,6380 (0,0144)	0,7286 (0,0241)	0,5171 (0,0242)	0,6366 (0,0252)	0,6366 (0,0323)	0,6364 (0,0134)	0,6365 (0,0144)
Glass	0,6773 (0,0134)	0,6773 (0,0145)	0,6740 (0,0124)	0,6474 (0,0121)	0,6740 (0,0131)	0,6703 (0,0125)	**0,6834** (0,0123)	0,5847 (0,0241)	0,5785 (0,0231)	0,5655 (0,0122)	0,5753 (0,0134)
Haberman	0,2266 (0,0514)	0,2393 (0,0846)	0,2078 (0,0463)	0,2583 (0,0834)	0,2293 (0,0486)	0,2978 (0,0730)	0,0770 (0,0694)	**0,3012** (0,1155)	0,1875 (0,0873)	0,2245 (0,0682)	0,2231 (0,0742)
ImageSeg	0,5171 (0,0321)	**0,6558** (0,0213)	0,5401 (0,0134)	0,6217 (0,0131)	0,5239 (0,0132)	0,5401 (0,0134)	0,4016 (0,0241)	0,5464 (0,0141)	0,5459 (0,0143)	0,5435 (0,0242)	0,5458 (0,0134)
Ionosphere	0,0483 (0,0329)	**0,0979** (0,0424)	0,0470 (0,0320)	0,0485 (0,0166)	0,0483 (0,0329)	0,0470 (0,0320)	0,0775 (0,0685)	0,0529 (0,0335)	0,0489 (0,0325)	0,0490 (0,0329)	0,0497 (0,0281)
Iris	0,6433 (0,0400)	0,5492 (0,0390)	0,5924 (0,0332)	0,5517 (0,0366)	0,60542 (0,0194)	0,4825 (0,0340)	0,6961 (0,0915)	0,6387 (0,0884)	**0,7488** (0,0511)	0,6344 (0,0597)	0,6799 (0,0567)
Parkinsons	0,2636 (0,0659)	0,2970 (0,0785)	0,2636 (0,0659)	0,2680 (0,0367)	0,2909 (0,0759)	0,2841 (0,0703)	0,6134 (0,2419)	0,5259 (0,2030)	**0,7359** (0,2231)	0,3714 (0,1246)	0,3624 (0,1230)
Pima	0,2173 (0,0740)	**0,2310** (0,0824)	0,2138 (0,0696)	0,2249 (0,0798)	0,2245 (0,0751)	0,2175 (0,0734)	0,1612 (0,1432)	0,2261 (0,1133)	0,2235 (0,1607)	0,1636 (0,0696)	0,1746 (0,0835)
QSAR	0,1790 (0,0351)	**0,2310** (0,0290)	0,1790 (0,0351)	0,1893 (0,0352)	0,1584 (0,0252)	0,2043 (0,0344)	0,0579 (0,0737)	0,2006 (0,0314)	0,1979 (0,0315)	0,1990 (0,0316)	0,1996 (0,0311)
Seeds	0,4555 (0,0736)	0,5759 (0,0205)	0,5745 (0,0263)	0,5738 (0,0208)	0,5856 (0,0219)	0,5495 (0,0264)	0,7232 (0,1169)	0,6640 (0,1031)	**0,8326** (0,1277)	0,4712 (0,0530)	0,4834 (0,0589)
Sonar	0,0557 (0,0134)	0,0528 (0,0103)	0,0557 (0,0241)	0,1208 (0,0131)	0,0574 (0,0313)	0,0474 (0,0364)	0,7445 (0,0231)	0,1424 (0,0312)	**0,8767** (0,0322)	0,1669 (0,0134)	0,1513 (0,0142)
StatlogV	0,5111 (0,0231)	0,5287 (0,0312)	0,5472 (0,0212)	0,5051 (0,0231)	0,5451 (0,0312)	0,5329 (0,0212)	0,6632 (0,0312)	**0,6814** (0,0212)	0,6542 (0,0213)	0,5607 (0,0241)	0,6101 (0,0231)
Thyroid	0,7965 (0,0495)	0,7748 (0,0799)	0,6752 (0,0428)	0,7729 (0,0811)	**0,8327** (0,0536)	0,7387 (0,0736)	0,4775 (0,1182)	0,6399 (0,0496)	0,6048 (0,1076)	0,4950 (0,0763)	0,5685 (0,1167)
Vertebral	0,2547 (0,1167)	0,3314 (0,1280)	0,3110 (0,1191)	0,3317 (0,1287)	0,3160 (0,1199)	0,3246 (0,1200)	0,2599 (0,1803)	**0,4583** (0,1695)	0,4148 (0,2418)	0,2499 (0,0838)	0,2741 (0,1035)
WDBC	0,2846 (0,0290)	0,2573 (0,0313)	0,2846 (0,0290)	0,2190 (0,0252)	0,3121 (0,0302)	0,2055 (0,0202)	0,4681 (0,0906)	0,2002 (0,0611)	**0,5238** (0,2150)	0,1866 (0,0799)	0,2151 (0,0836)
Wine	0,5814 (0,1286)	0,5313 (0,1175)	0,5814 (0,1286)	0,3511 (0,0000)	0,5814 (0,1286)	0,4993 (0,1228)	0,7488 (0,1520)	0,5880 (0,1503)	**0,7614** (0,1316)	0,4994 (0,0876)	0,5143 (0,0845)
WineQlt	0,6328 (0,0312)	0,6326 (0,0312)	0,6404 (0,0214)	0,6220 (0,0324)	0,6291 (0,0352)	0,6332 (0,0351)	**0,6549** (0,0311)	0,6513 (0,0251)	0,6478 (0,0144)	0,6238 (0,0141)	0,6430 (0,0411)
Yeast	0,7001 (0,0312)	0,7271 (0,0311)	0,7271 (0,0312)	0,7151 (0,0124)	0,6105 (0,0142)	0,7043 (0,0312)	**0,7740** (0,0142)	0,7394 (0,0241)	0,7394 (0,0242)	0,7029 (0,0252)	0,7030 (0,0221)

Tabela 4.15: Média e desvio padrão (entre parênteses) do índice FR (maiores valores em negrito) para métodos de agrupamento e conjuntos de dados da UCI.

o método MFCM foi superior 6 vezes, o método MFCM-G 3 vezes e o método MFCM-D 8 vezes. É importante observar também que os métodos multivariados baseados na abordagem possibilística MPFCM e MPFCM-D não conseguiram ser superiores em relação a todos os outros para os dados reais analisados. Nota-se também que a versão ponderada MPFCM-D foi tão bom ou superior a sua versão não ponderada MPFCM na maioria dos conjuntos de dados reais.

Com o objetivo de compreender o desempenho dos métodos de agrupamento de uma forma mais global em relação aos conjuntos de dados reais, a Tabela 4.16 apresenta, para cada conjunto, a posição de cada método segundo a sua média do índice FR obtida de forma

Conjunto	GK	FCM	FCM-C[1]	PFCM	FCM-C[2]	W-FCM	MFCM	MFCM-G	MFCM-D	MPFCM	MPFCM-D
Abalone	6	8	7	10	4	9	1	11	2	5	3
Banknote	8	2	8	5	7	3	1	4	11	10	6
Blood	1	4	5	3	6	2	11	7	10	9	8
Breast	4	7	5	9	3	10	2	11	1	8	6
BreastTissue	7	6	9	11	7	10	4	5	1	3	2
Climate	9	6	9	7	11	8	1	3	5	4	2
Ecoli	3	1	5	2	5	4	11	7	7	10	9
Glass	2	2	4	7	4	6	1	8	9	11	10
Haberman	6	4	9	3	5	2	11	1	10	7	8
ImageSeg	10	1	7	2	9	7	11	3	4	6	5
Ionosphere	8	1	10	7	8	10	2	3	6	4	4
Iris	4	9	7	8	11	10	2	5	1	6	3
Parkinsons	10	6	10	9	7	8	2	3	1	4	5
Pima	7	1	8	3	4	6	11	2	5	10	9
QSAR	8	1	8	7	10	2	11	3	6	4	4
Seeds	11	5	6	7	4	8	2	3	1	10	9
Sonar	8	10	8	6	7	11	2	5	1	4	3
StatlogV	10	9	6	11	7	8	2	1	3	5	4
Thyroid	2	3	6	4	1	5	11	7	8	10	9
Vertebral	10	4	7	3	6	5	9	1	2	11	8
WDBC	4	6	4	7	3	9	2	10	1	11	8
Wine	4	7	4	11	4	10	2	3	1	9	8
WineQlt	7	8	5	11	9	6	1	2	3	10	4
Yeast	4	10	4	6	11	7	1	2	2	9	8
Média	6,3750	5,0416	6,7083	6,6250	6,3750	6,9166	4,7500	4,5833	4,2083	7,5000	6,0416
DP	2,8840	3,0066	1,9035	2,9413	2,7358	2,7676	4,3036	3,0127	3,3289	2,7386	2,5245

Tabela 4.16: Ordenação dos métodos de agrupamento quanto às maiores médias do índice FR para os conjuntos de dados reais da UCI.

decrescente. Isto é, o método correspondente a maior média de FR dentre todos os métodos de um determinado conjunto receberá 1; por outro lado, o que obteve a menor média receberá 11. As duas últimas linhas da tebela apresentam a média e o desvio padrão das posições para cada método de agrupamento. Segundo a tabela, o método multivariado com ponderação na distância MFCM-D obteve a menor média, o método com podenderação no grau MFCM-G obteve a segunda menor média e o sem poderação MFCM a terceira dentre todos os métodos avaliados. No contexto dos métodos baseados na abordagem possibilística, em média, MPFCM e MPFCM-D foram inferiores aos métodos MFCM, MFCM-G e MFCM-D. Além disso, o método multivariado ponderado baseado na abordagem possibilística MPFCM-D obteve uma média menor que a sua versão não ponderada e que o tradicional PFCM.

4.4 Interpretação de Partições e Grupos Difusos: Conjunto de Dados Abalone

Esta seção apresenta a utilidade dos índices de interpretação através dos resultados obtidos com a aplicação dos métodos multivariados MFCM, MFCM-G, MFCM-D, MPFCM e MPFCM-D para o conjunto de dados Abalone. Este conjunto diz respeito à análise de moluscos

gastrópodes através de medidas e cortes na sua concha e são classificado de acordo com o sexo: *Male* (`M`), *Female* (`F`) e *Infant* (`I`). Para melhor compreender este conjunto, a Tabela 4.17 enumera as suas variáveis segundo o seu nome, abreviação, tipo, unidade de medida e descrição.

Nome	Abreviação	Tipo	Uni. Medida	Descrição
Length	`Lh`	Contínua	milímetro	Maior medida da concha
Diameter	`Dm`	Contínua	milímetro	Perpendicular ao comprimento
Height	`Ht`	Contínua	milímetro	Com a carne na concha
Whole weight	`We`	Contínua	grama	Considerando todo o animal
Shucked weight	`Sd`	Contínua	grama	Peso da carne
Viscera weight	`Va`	Contínua	grama	Peso do intestino (após sangrar)
Shell weight	`Sl`	Contínua	grama	Depois de ter sido seco
Rings	`Rs`	Inteira	-	+1,5 dá a idade em anos

Tabela 4.17: Descrição das variáveis para o conjunto de dados Abalone.

Tabela 4.18 mostra o índice de heterogeneidade global R para os três métodos. Estes valores estão perto de 1. Assim, os padrões pertencentes aos grupos são bem representados pelos protótipos do grupo correspondentes. Outro ponto importante que pode ser notado pela tabela é a superioridade do valores deste índice correspondente aos métodos que utilizam ponderação (MFCM-G, MFCM-D e MPFCM-D).

R^1	$0,7131$
R^2	$0,8090$
R^3	$0,8029$
R^4	$0,7534$
R^5	$0,8011$

Tabela 4.18: Índice de heterogeneidade global para os métodos multivariados.

Uma comparação entre os índices $COR(j)$ e $CTR(j)$ para cada variável j do conjunto de dados Abalone e método é mostrado na Tabela 4.19. Em relação ao índice de $COR(j)$ para os métodos MFCM e MFCM-D, as variáveis `We` e `Sd` não são tão úteis para produzir uma alta qualidade de partição. Por outro lado, para MFCM-G, as variáveis `Lh`, `Dm`, `Ht` e `We` podem produzir boa qualidade de partição. Para os métodos MPFCM e MPFCM-D, a variável `Sd` é útil para produzir uma alta qualidade de partição. Quanto ao índice $CTR(j)$, a variável `Rs` tem uma contribuição importante para a separação dos representantes dos grupos para o método MFCM, MPFCM e MPFCM-D. Por outro lado, `We` e `Ht` são as variáveis com maior poder discriminante para MFCM-G e MFCM-D, respectivamente.

Tabela 4.20 e 4.21 mostram os índices $T^1(i)$, $B^1(i)$ e $J^1(i)$ para cada grupo i. A partir

Método	Medida	Variável							
		Lh	Dm	Ht	We	Sd	Va	Sl	Rs
MFCM	COR^1	0,2946	**0,9601**	0,9292	0,3167	0,1631	0,7956	0,8047	0,9156
	CTR^1	0,0000	0,0009	0,0001	0,1043	0,0015	0,0002	0,0004	**0,8926**
MFCM-G	COR^2	0,6580	0,6293	0,5554	**0,6820**	0,0086	0,0667	0,0945	0,1439
	CTR^2	0,2741	0,2517	0,1353	**0,2998**	0,0012	0,0088	0,0121	0,0171
MFCM-D	COR^3	0,2719	**0,9618**	0,9312	0,3139	0,1556	0,8083	0,8133	0,9071
	CTR^3	0,0063	**0,3696**	0,2479	0,0117	0,0054	0,0666	0,0713	0,2211
MPFCM	COR^4	0,2421	**0,8780**	0,8431	0,2452	0,1373	0,6835	0,7845	0,8462
	CTR^4	0,0035	0,0014	0,0295	0,1381	0,0184	0,0184	0,0038	**0,7869**
MPFCM-D	COR^5	0,2857	0,8572	**0,9385**	0,2863	0,1535	0,7846	0,7639	0,8964
	CTR^5	0,0184	**0,2857**	0,2356	0,0284	0,0074	0,0576	0,0938	0,2740

Tabela 4.19: Índice de heterogeneidade global e inter-grupo de acordo com cada variável e método.

desses valores, é possível concluir que grupo F é o mais próximo do vetor médio global para MFCM e MFCM-G, enquanto que para os métodos MFCM-D, MPFCM e MPFCM-D é o grupo I. Para os métodos MFCM, MFCM-G e MPFCM, o grupo F é o mais distante, enquanto para MFCM-D e MPFCm-D é o grupo M. Para os métodos MFCM e MFCM-G, o grupo I é o mais homogêneo; para o MFCM-D e MPFCM-D, o mais homogêneo é o grupo F e para o MPFCM, o mais homogêneo é o grupo M.

Grupo	MFCM			MFCM-G			MFCM-D		
	$T^1(i)$	$B^1(i)$	$J^1(i)$	$T^2(i)$	$B^2(i)$	$J^2(i)$	$T^3(i)$	$B^3(i)$	$J^3(i)$
M	**0,3278**	0,4290	0,3204	**0,3223**	0,4180	0,3126	0,4643	0,4520	0,3333
F	0,3424	**0,1841**	0,3754	0,3476	**0,1930**	0,3986	0,3506	0,4007	**0,3272**
I	0,3298	0,3869	**0,3042**	0,3301	0,3890	**0,2888**	**0,1851**	**0,1474**	0,3396

Tabela 4.20: Contribuição relativa de grupos de medidas baseadas na soma de quadrados para os métodos MFCM, MFCM-G e MFCM-D.

Grupo	MPFCM			MPFCM-D		
	$T^4(i)$	$B^4(i)$	$J^4(i)$	$T^5(i)$	$B^5(i)$	$J^5(i)$
M	0,3284	0,4183	**0,3028**	0,4482	0,4448	0,3463
F	0,3683	**0,1948**	0,3673	0,3736	0,4218	**0,3184**
I	**0,3033**	0,3869	0,3299	**0,1782**	**0,1334**	0,3353

Tabela 4.21: Contribuição relativa de grupos de medidas baseadas na soma de quadrados para os métodos MPFCM e MPFCM-D.

Tabelas 4.22, 4.23, 4.24, 4.25 e 4.26 mostram os índices de heterogeneidade decompostos de acordo com cada variável para os métodos MFCM, MFCM-G MFCM-D, MPFCM e MPFCM-D. A partir da Tabela 4.22, variáveis `Sl`, `Lh` e `Dm` tem um comportamento relativamente homogêneo intra-grupos `M`, `F` e `I`, respectivamente. Variável `Rs` é muito importante para a separação dos grupos.

Variável	Grupo								
	M			F			I		
	COR	CTR	CE	COR	CTR	CE	COR	CTR	CE
Lh	0,0438	0,0000	0,0000	**0,2351**	0,0001	0,0000	0,0157	0,0000	0,0000
Dm	0,3862	0,0008	0,0004	0,2141	0,0011	0,0002	**0,3598**	0,0009	0,0003
Ht	0,3888	0,0001	0,0000	0,1888	0,0001	0,0000	0,3517	0,0001	0,0000
We	0,1763	0,1354	0,0581	0,0039	0,0069	0,0013	0,1366	0,1163	0,0450
Sd	0,0658	0,0014	0,0006	0,0084	0,0004	0,0001	0,0889	0,0021	0,0008
Va	0,3924	0,0002	0,0001	0,0638	0,0001	0,0000	0,3395	0,0002	0,0001
Sl	**0,3926**	0,0004	0,0002	0,0735	0,0002	0,0000	0,3387	0,0004	0,0001
Rs	0,3792	**0,8617**	**0,3697**	0,1872	**0,9911**	**0,1825**	0,3492	**0,8800**	**0,3405**

Tabela 4.22: Índices de interpretação por grupo e variável para o método MFCM.

De acordo com a Tabela 4.23, as variáveis `Dm`, `Ht` e `We` são as que mais contribuem para a homogeneidade dentro dos grupos `M`, `F` e `I`, respectivamente. Por outro lado, a variável `Dm` contribui para separar o grupo `M` e variável `We` contribui para separar os grupos `F` e `I`.

Variável	Grupo								
	M			F			I		
	COR	CTR	CE	COR	CTR	CE	COR	CTR	CE
Lh	0,3201	0,2951	0,1334	0,2951	0,3068	0,1229	0,0427	0,1208	0,0178
Dm	**0,5470**	**0,4840**	**0,2188**	0,0811	0,0809	0,0324	0,0013	0,0035	0,0005
Ht	0,0526	0,0283	0,0128	**0,3525**	0,2143	0,0859	0,1503	0,2484	0,0366
We	0,1636	0,1591	0,0719	0,3286	**0,3605**	**0,1444**	**0,1898**	**0,5661**	**0,0834**
Sd	0,0027	0,0008	0,0004	0,0032	0,0011	0,0004	0,0028	0,0026	0,0004
Va	0,0212	0,0062	0,0028	0,0200	0,0066	0,0026	0,0255	0,0228	0,0034
Sl	0,0315	0,0089	0,0040	0,0315	0,0101	0,0040	0,0315	0,0274	0,0040
Rs	0,0666	0,0175	0,0079	0,0666	0,0198	0,0079	0,0106	0,0085	0,0013

Tabela 4.23: Índices de interpretação por grupo e variável para o método MFCM-G.

Os valores da Tabela 4.24 indicam que as variáveis `Va`, `Lh` e `Ht` apresentam um comportamento relativamente homogêneo dentro do grupo `M`, `F` e `I`, respectivamente. Por outro lado, variável `Dm` é a mais importante para permitir a separação entre os grupos.

Variável	Grupo								
	M			F			I		
	COR	CTR	CE	COR	CTR	CE	COR	CTR	CE
Lh	0,0418	0,0023	0,0010	**0,2149**	0,0257	0,0050	0,0153	0,0009	0,0004
Dm	0,3848	**0,3538**	**0,1479**	0,2042	0,4067	**0,0785**	**0,3728**	**0,3682**	**0,1432**
Ht	0,3845	0,2449	0,1024	0,1892	**0,2610**	0,0504	0,3574	0,2446	0,0952
We	0,1770	0,0158	0,0066	0,0042	0,0008	0,0002	0,1327	0,0128	0,0050
Sd	0,0651	0,0054	0,0023	0,0109	0,0020	0,0004	0,0796	0,0071	0,0028
Va	**0,3949**	0,0779	0,0325	0,0639	0,0273	0,0053	0,3494	0,0740	0,0288
Sl	0,3925	0,0823	0,0344	0,0735	0,0334	0,0064	0,3473	0,0783	0,0305
Rs	0,3730	0,2175	0,0909	0,1924	0,2431	0,0469	0,3416	0,2141	0,0833

Tabela 4.24: Índices de interpretação por grupo e variável para o método MFCM-D.

Segundo a Tabela 4.25, as variáveis Sl, Lh e Va apresentam um comportamento relativamente homogêneo dentro do grupo M, F e I, respectivamente. No que diz respeito à contribuição relativa de cada variável para a medida inter-grupo (isto é, a separação dos grupos), a variável Rs mostrou ser a mais importante para este método.

Variável	Grupo								
	M			F			I		
	COR	CTR	CE	COR	CTR	CE	COR	CTR	CE
Lh	0,1849	0,0019	0,0015	**0,2274**	0,0038	0,0018	0,0284	0,0019	0,0029
Dm	0,3649	0,0019	0,0008	0,2038	0,0010	0,0019	**0,3329**	0,0290	0,0007
Ht	0,3563	0,0008	0,0360	0,1937	0,0037	0,0099	0,3129	0,0009	0,0029
We	0,1847	0,1295	0,0573	0,0184	0,0070	0,0039	0,1294	0,1492	0,0674
Sd	0,0573	0,0282	0,0001	0,0148	0,0765	0,0013	0,0847	0,0390	0,0299
Va	0,3673	0,0019	0,0038	0,0737	0,0008	0,0039	0,3285	0,0029	0,0039
Sl	0,3684	0,0009	0,0023	0,0847	0,0002	0,0001	0,3194	0,0030	0,0078
Rs	**0,4839**	**0,8349**	**0,3165**	0,1938	**0,2185**	**0,1720**	0,3048	**0,7741**	**0,2714**

Tabela 4.25: Índices de interpretação por grupo e variável para o método MPFCM.

Por fim, para o método MPFCM-D e segundo a Tabela 4.26, nota-se que a variável Dm tem um comportamento homogêneo nos grupos M e I e a variável Ht tem no grupo F. A variável Dm também mostrou ser importante para a separação dos grupos M e I. Por outro lado, a que mais contribuiu para a separabilidade do grupo F foi a variável Rs.

Variável	Grupo								
	M			F			I		
	COR	CTR	CE	COR	CTR	CE	COR	CTR	CE
Lh	0,0329	0,0047	0,0038	0,1928	0,0184	0,0038	0,0148	0,0018	0,0017
Dm	0,3747	0,3375	**0,1395**	0,1845	**0,3950**	0,0673	**0,3634**	**0,3608**	**0,1358**
Ht	0,3537	0,2284	0,1194	**0,2948**	0,2639	0,0485	0,3476	0,2437	0,0797
We	0,1683	0,0127	0,0088	0,0079	0,0018	0,0038	0,1284	0,0184	0,0048
Sd	0,0638	0,0076	0,0068	0,0284	0,0039	0,0018	0,0737	0,0068	0,0047
Va	**0,3837**	0,0860	0,0486	0,0385	0,0294	0,0062	0,3379	0,0684	0,0223
Sl	**0,3827**	0,0590	0,0323	0,0878	0,0294	0,0089	0,3284	0,0639	0,0348
Rs	0,3284	**0,3453**	0,0856	0,1328	0,3423	**0,2815**	0,3562	0,2563	0,0515

Tabela 4.26: Índices de interpretação por grupo e variável para o método MPFCM-D.

5

APLICAÇÃO EM DADOS BIOLÓGICOS

A coisa mais indispensável a um homem é reconhecer o uso que deve fazer do seu próprio conhecimento.

—PLATÃO

5.1 Introdução a Bioinformática

Nesta seção, serão apresentados conceitos de Biologia e Bioinformática utilizados ao decorrer desse capítulo para um melhor entendimento dos experimentos realizados. **Bioinformática** (ou também chamada de **Biologia Computacional**) refere-se à resolução de problemas da Biologia através da utilização de técnicas e ferramentas de computação (Baldi e Brunak, 2001). Exemplos onde a computação pode ser aplicada é a comparação de sequencias (DNA, RNA e proteínas), montagem de fragmentos, reconhecimento de genes, identificação e análise das expressões gênicas e determinação da estrutura de proteínas (Setúbal e Meidanis, 1997; De Souto et al., 2003). Uma dificuldade encontrada em muitos problemas biológicas é a extração de informação a partir de uma grande quantidade de dados. Nesse contexto, técnicas de Aprendizado de Máquina são empregadas por sua capacidade de aprender automaticamente (Mitchell , 1997).

Um dos principais objetos de estudo da Bioinformática é o **genoma** dos organismos, o qual constitui o conjunto de suas informações genéticas. Desta forma, o genoma pode ser compreendido como uma coleção de fatores responsáveis pela hereditariedade de um organismo vivo. Esse fatores são denominados **genes** e são responsáveis por codificar toda a informação genética (Slack, 2014). Os **cromossomos** são estruturas biológicas com funções específicas presentes em todas as células de um organismo e contém vários genes, sendo assim seus

portadores. Estudos sobre cromossomos mostraram que eles são constituídos por sequências de moléculas de **Ácido Desoxirribonucléico** (do inglês, *Deoxyribonucleic Acid* — DNA) (Alberts et al., 2002). A Figura 5.1 esquematiza a estrutura de um cromossomo e sua localização no núcleo celular.

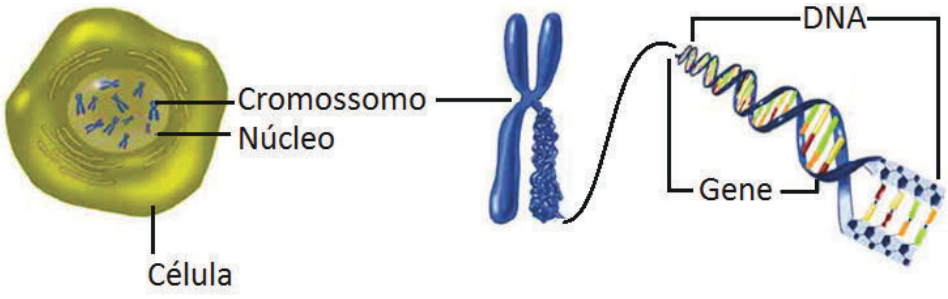

Figura 5.1: Relação entre célula, cromossomo, DNA e gene.

Uma molécula de DNA é formada por duas fitas anti-paralelas entrelaçadas em forma de dupla hélice, onde cada fita é composta por uma sequência de nucleotídeos (ou bases), as quais podem ser: Adenina (A), Guanina (G), Citosina (C) e Timina (T) (Butler, 2005). A Figura 5.2 exemplifica uma estrutura de DNA com associações entre as bases. Um DNA pode conter diversos genes, que por sua vez realizam uma importante função biológica: codificam protéinas. As **proteínas**, formadas por sequências de aminoácidos, são essenciais para um ser vivo porque exercem funções relevantes tais como estruturais, regulatórias ou catalizadoras. O processo pelo qual as sequências de nucleotídeos dos genes são interpretadas na produção de proteínas é denominado **expressão gênica** (Ji et al. , 2014).

Figura 5.2: Exemplo de uma estrutura de DNA com suas ligações entre bases.

A análise de expressão gênica é importante para a Biologia porque pode ser uma fonte de informação a respeito do comportamento de células, já que alterações fisiológicas de um organismo estão normalmente relacionadas a alterações na expressão gênica (Alberts et al., 2002). Diversas técnicas podem ser utilizadas para fazer a análise de expressão gênica, tais

como MPSS (*Massively Parallel Signature Sequence technology*), SAGE (*Serial Analysis of Gene Expression*) ou *microarray* de DNA (Brenner et al., 2000). Dentre estas técnicas, uma das mais utilizadas é a *microarray* onde medidas de níveis de expressões gênicas referentes a uma ampla variedade de genes pode ser armazenada em uma única superfície sólida.

Para fazer o armazenamento de expressões gênicas em *microarray*, amostras biológicas são dispostas em superfícies adequadamente preparadas, que são divididas em milhares de pontos. Cada ponto contém milhares de sondas de cDNA (DNA sintetizado a partir de uma molécula de RNA mensageiro) que foram projetadas para hibridizar com o mRNA (RNA mensageiro, responsável pela transferência de informação) de um certo gene. As moléculas de cDNA produzidas são, em geral, marcadas com rótulos fluorescentes verdes (Cy3), enquanto que as moléculas de mRNA são marcadas com rótulos vermelhos (Cy5). Em seguida, a lâmina é escaneada utilizando-se *lasers*, produzindo como resultado uma imagem com as intensidades luminosas dos rótulos de todos os pontos. A partir dessa lâmina final, uma análise computacional é realizada para gerar dados, relacionando genes, amostras e variáveis quantitativas (Duggan et al., 1999). A Figura 5.3 mostra o processo de criação de um *microarray*.

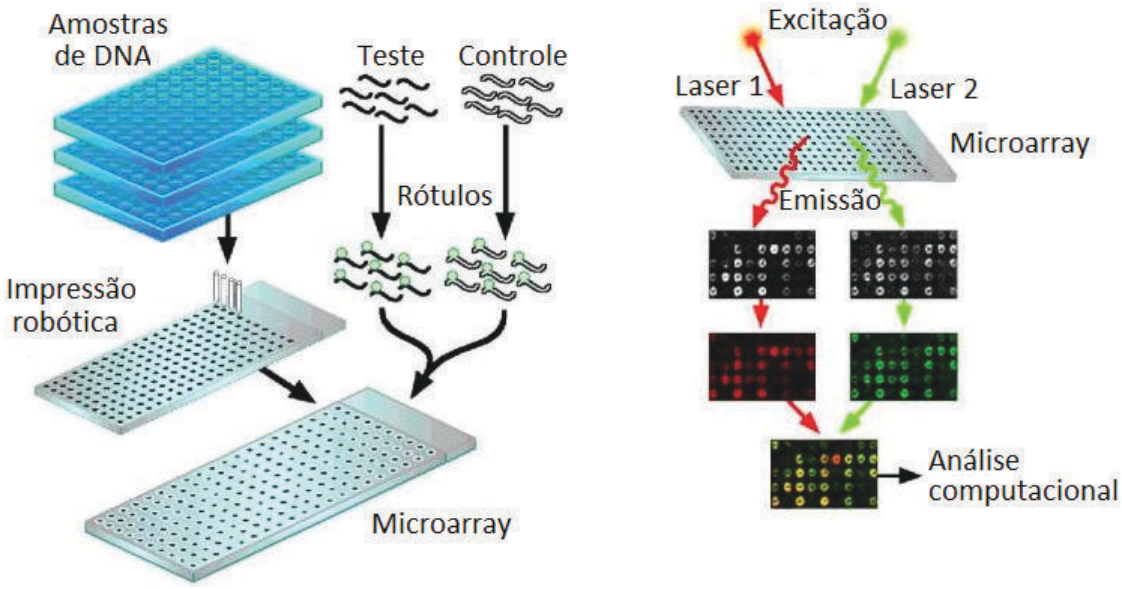

Figura 5.3: Constituição de um *microarray* a partir de amostras de DNA e utilização de *lasers* para a excitação dos rótulos fluorescentes.

5.2 Motivação

Com o crescente interesse em entender, processar e resumir dados, muitos domínios de aplicações tais como Reconhecimento de Padrões, Aprendizado de Máquina, Mineração de Dados, Visão Computacional e Bioinformática têm desenvolvido técnicas e métodos buscando automatizar a extração de informação dos dados. No caso de Bioinformática, dados são essenciais para a pesquisa biológica. Uma metodologia bastante conhecida neste tipo de pesquisa é o *microarray*, o qual armazena grande quantidade de material biológico. Experimentos com *microarray* são capazes de monitorar simultaneamente níveis de milhares de expressões gênicas em um organismo (Bhaskar, 2006; Dembélé, 2013).

Inicialmente, aplicações de *microarray* foram feitas para estudar o comportamento de doenças humanas servindo como uma ferramenta para diagnóstico médico (Golub et al., 1999; Alizadeh et al., 2000; Khan et al., 2001). Um *microarray* pode ser classificado em várias categorias dependendo do material biológico que ele contém. Por exemplo, *microarray* de composto químico, de proteína ou de DNA, sendo este último o mais utilizado. Recentemente, há um número crescente de banco de dados com informações complexas de expressões gênicas (Choi, 2003). Uma tarefa desafiadora para o especialista é como extrair, comparar e integrar informações a partir uma grande quantidade de dados (Choi, 2003). Uma forma de extrair informação destes dados é a utilização de técnicas e métodos de Aprendizado de Máquina (Bhaskar, 2006).

Atualmente, agrupamento do tipo particional é o mais frequentemente utilizado para obter informações biológicas por meio de grupos baseando-se na similaridade entre os dados (Boeva, 2013; Anderson, 1986; Witten et al., 2011). Mais especificamente, o método particional rígido mais utilizado é o *K-Means* (Ichikawa e Morishita, 2014; MacKay, 2003; Blum e Chawla, 2001; Breve et al., 2012). É conhecido por ser fácil de programar e rápido na execução. Entretanto, ele supõe que todos os grupos possuem formas esféricas, o que o torna não adequado em muitos problemas reais onde os grupos possuem diferentes dispersões entre as variáveis. Além disso, por ser do tipo rígido, ele não é adequado em problemas em que os grupos são sobrepostos. Uma vantagem do uso de métodos difusos é a possibilidade de interpretar o valor do grau de pertinência: quão semelhante um elemento é de um dado grupo. Entretanto, um ponto em aberto em Bioinformática é como identificar qual variável contribuiu para o grau de semelhança entre o elemento e o grupo. Desta forma, a proposta é utilizar a abordagem multivariada.

5.3 Objetivo

O método *Multivariate Fuzzy C-Means* (MFCM) foi proposto buscando atender dois pontos: (1) obtenção de uma melhor qualidade de agrupamento obtida pelos métodos; (2) capacidade de interpretar o grau de pertinência de cada objeto para um determinado grupo de acordo com cada variável. Até então neste trabalho, foi explorado apenas o primeiro ponto mostrando que os métodos multivariados são superiores em determinados cenários. Por outro lado, nada foi explorado a respeito da interpretação dos graus de pertinência multivariados. Portanto, o objetivo deste capítulo é apresentar uma aplicação dos métodos multivariados através dos seus graus utilizando conjunto de dados biológicos.

Existem bancos de dados biológicos que podem ser usados para avaliar o desempenho dos métodos. São bibliotecas de informações científicas coletadas através de experimentos, publicados na literatura e armazenados usando técnicas computacionais (Attwood et al., 2011). Esses bancos são importantes ferramentas para ajudar cientistas na análise e explicação dos fenômenos biológicos mediante a estrutura de moléculas e suas interações. Dependendo do tipo de material biológico que armazenam, os bancos de dados podem ser classificados em genômicos, protéicos, metabólicos, *microarray* de expressões gênicas e filogênicos (Altman, 2004). O *International Nucleotide Sequence Database Collaboration* (INSDC), por exemplo, é um esforço conjunto para divulgar três bancos de dados de sequencias de DNA e RNA: *DNA Data Bank of Japan* (Japão), GenBank (mantida pela *National Center for Biotechnology Information* - NCBI, Estados Unidos) e *European Nucleotide Archive* (Reino Unido) (Karsch-Mizrachi et al., 2012). Neste trabalho, as informações sobre os genes foram coletadas utilizando dados disponibilizados pelo site da NCBI.

5.4 Uma Estratégia para Interpretar Graus de Pertinência Multivariados

A suposição subjacente à estratégia proposta neste trabalho é a seguinte. O grau de pertinência de uma dada variável para um determinado objeto deve ser maior no grupo em que tal variável é mais discriminativa. Ou seja, ele está mais perto do protótipo em relação a uma variável. Neste contexto, também assume-se que quanto maior o grau de pertinência de uma variável para um grupo, mais relevante esta variável é para caracterizá-lo. Com base nisto, propõe-se uma estratégia simples (Algoritmo 12), que consiste no cálculo da média (a_{ij}) do grau de pertinência multivariado levando em conta todos os objetos. O algoritmo a seguir explica como a relevância de uma determinada variável j em um grupo i é calculado:

Algoritmo 12 Relevância$(i, j, \mathbf{U})$

Requer: Grupo i, variável j e matriz de graus de pertinência multivariados $\mathbf{U}$;
1: Inicialize $a_{ij} = 0$;
2: **Para** cada objeto k do conjunto $(k = 1, \ldots, n)$ **Faça**
3: Pegue o grau de pertinência multivariado correspondente ao grupo C_i e variável j (u_{ijk});
4: $a_{ij} = a_{ij} + u_{ijk}$.
5: **Fim Para**
6: $a_{ij} = a_{ij}/n$;
Retorne a_{ij}.

Para fins pedagógicos, considere um exemplo muito simples da utilização do Algoritmo 12. Suponha o conjunto de dados sintético ilustrada na Figura 5.4. Tal conjunto de dados consiste de quatro classes bem separadas, onde as classes 1 e 3 são melhores descritas pela variável 2, enquanto que as classes 2 e 4 são melhores descritas pela variável 1. Cada classe é desenhada de acordo com uma distribuição normal bi-variada, cujos componentes são variáveis independentes. Tabela 5.1 mostra os parâmetros das classes para este conjunto de dados sintético.

Classe	μ_1	μ_2	σ_1^2	σ_2^2	#Obj
1	40	7	100	4	100
2	70	37	4	100	100
3	40	70	100	4	100
4	10	37	4	100	100

Tabela 5.1: Parâmetros para o conjunto de dados sintético.

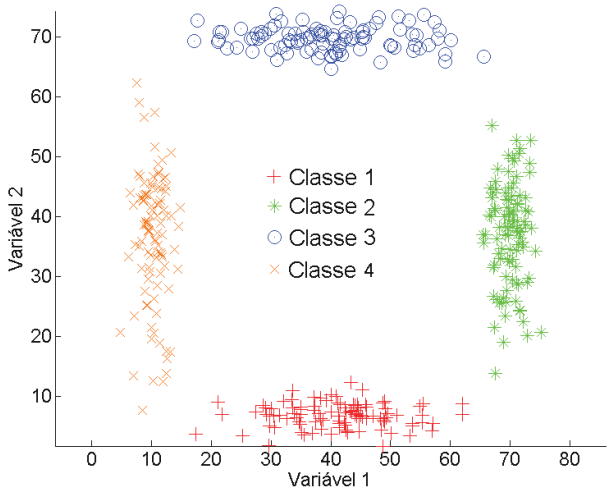

Figura 5.4: Dispersão do conjunto de dados sintético.

Método	Var.	C1	C2	C3	C4
MFCM	1	0,0652 (0,0152)	**0,1793** (0,0102)	0,0732 (0,0093)	**0,1775** (0,0112)
	2	**0,1864** (0,0136)	0,0701 (0,0081)	**0,1708** (0,0144)	0,0686 (0,0145)
MFCM-G	1	0,0387 (0,0111)	**0,1996** (0,0084)	0,0376 (0,0123)	**0,2084** (0,0164)
	2	**0,2071** (0,0101)	0,0502 (0,0122)	**0,2174** (0,0114)	0,0411 (0,0127)
MFCM-D	1	0,0000 (0,0000)	**0,2136** (0,0128)	0,0000 (0,0000)	**0,2546** (0,0102)
	2	**0,2696** (0,0145)	0,0000 (0,0000)	**0,2607** (0,0125)	0,0026 (0,0023)
MPFCM	1	0,0702 (0,0013)	**0,2101** (0,0121)	0,0661 (0,0131)	**0,1955** (0,0136)
	2	**0,1814** (0,0151)	0,0446 (0,0014)	**0,1783** (0,0131)	0,0535 (0,0112)
MPFCM-D	1	0,0378 (0,0013)	**0,2033** (0,0121)	0,0413 (0,0131)	**0,2162** (0,0142)
	2	**0,2139** (0,0135)	0,0399 (0,0112)	**0,2074** (0,0124)	0,0399 (0,0012)

Tabela 5.2: Média e desvio padrão (em parênteses) de a_{ij} dos métodos multivariados e conjunto de dados sintético.

A fim de analisar os graus de pertinência multivariados, os métodos multivariados foram executados com este conjunto de dados utilizando-se o parâmetro $m = 2$. Em termos de metodologia, os experimentos foram realizados utilizando Monte Carlo. Mais especificamente, 100 replicações foram realizadas de acordo com os parâmetros na Tabela 5.1. Para cada replicação, cada método foi executado 50 vezes com diferentes inicializações até a sua convergência. O melhor resultado de acordo com a função objetivo foi selecionado. Para validar a partição difusa dos métodos em contraste com a partição do conjunto de dados, calculou-se o índice FR como apresentado na Equação 4.1.

Após a execução das 100 replicações, a média e o desvio padrão deste índice foram calculados. Estes valores são: $0,8213$ e $0,0231$ (MFCM), $0,8769$ e $0,0134$ (MFCM-G), $0,8430$ e $0,0141$ (MFCM-D), $0,8072$ e $0,0134$ (MPFCM), $0,8611$ e $0,0141$ (MPFCM-D). Assim, em média, os métodos conseguiram produzir partições próximas da verdadeira. A fim de verificar se a suposição de que maior o grau de pertinência, maior a relevância para caracterizar o grupo, o algoritmo 12 foi executado para cada partição difusa. O resultado final é a média de 100 valores de a_{ij}. Tabela 5.2 mostra a média e desvio padrão (entre parênteses) de a_{ij} obtido pelos métodos multivariados para o conjunto de dados sintético.

De acordo os resultados dos métodos multivariados apresentados na Tabela 5.2, em

relação à variável 2, a média de a_{ij} para os grupos 1 e 3 é maior do que para os grupos 2 e 4. Em contraste, a média de a_{ij} para os grupos 2 e 4 é maior do que para os grupos 1 e 3 com respeito à variável 1. Portanto, como esperado, os grupos 1 e 3 podem ser melhor descritos pela variável 2, enquanto os grupos 2 e 4 podem ser melhor descritos pela variável 1.

5.5 Aplicação

Oito conjuntos de dados de expressões gênicas, que são referência em estudos presentes na litetura de bioinformática (ver De Souto et al. (2008)), foram utilizadas. A Tabela 5.3 mostra uma breve descrição desses conjuntos segundo o número de amostras, o número de classes, a quantidade de amostras em cada classe e o número de genes que descrevem cada amostra. Estes conjuntos de dados foram escolhidos de forma a ter representantes das tecnologias mais utilizadas para obtenção de dados de expressão: chips Affymetrix de único canal (ou simplesmente Affymetrix) e cDNA de duplo canal (ou cDNA). Além disso, para ter alguma diversidade entre os conjuntos de dados escolhidos há alguns que, no estudo em De Souto et al. (2008), as partições geradas pelo K-Means apresentaram grande, médio e baixo valor de CR. Os conjuntos de dados analisados neste estudo estão disponíveis em `http://bioinformatics.rutgers.edu/Supplements/CompCancer/`.

	Conjunto	# Amostras	# Classes	Cardinalidade	# Genes
Affymetrix	`Chowdary`	104	2	62, 42	182
	`Dryskjot`	40	3	9, 21, 11	1203
	`Gordon`	181	2	31, 150	1626
	`Nutt2`	28	2	14, 14	1070
	`Shipp`	77	2	58, 19	798
cDNA	`Alizadeh1`	42	2	21, 21	1095
	`Chen`	180	2	104, 76	85
	`Khan`	83	4	29, 11, 18, 25	1069

Tabela 5.3: Descrição de conjuntos de dados de expressão de genes.

Os experimentos foram realizados com os métodos presentes na literatura apresentados no capítulo anterior e os multivariados propostos neste trabalho. Em termos de metodologia, no âmbito de um experimento de *Booststrap*, 50 reamostragens da cada conjunto de dados foram construídas. Para cada replicação, os métodos de agrupamento foram aplicadas aos conjuntos de dados com 50 inicializações aleatórias e eles foram executados até a sua convergência. Em seguida, o melhor resultado de acordo com a função de objetivo foi selecionado e o índice FR foi calculado comparando a partição difusa obtida por cada método de agrupamento e a verdadeira

partição. Após a execução das 50 reamostragens, a média e o desvio padrão do FR foram calculados. Apêndice B apresenta os testes de hipótese relativos aos resultados apresentados.

Tabela 5.4 a média e desvio padrão (entre parênteses) do índice FR obtido pelos métodos para conjuntos de dados de genes. O maior valor médio da FR para cada conjunto de dados está em negrito.

	Conjunto	GK	FCM	FCM-C[1]	PFCM	FCM-C[2]	W-FCM	MFCM	MFCM-G	MFCM-D	MPFCM	MPFCM-D
Affymetrix	Chowdary	0,494 (0,013)	0,348 (0,021)	0,486 (0,013)	0,347 (0,011)	0,250 (0,017)	0,298 (0,021)	**0,556** (0,025)	0,358 (0,019)	0,185 (0,014)	0,484 (0,011)	0,488 (0,010)
	Dryskjot	0,215 (0,011)	0,178 (0,012)	0,201 (0,015)	0,175 (0,021)	0,147 (0,018)	0,215 (0,014)	**0,243** (0,009)	0,080 (0,011)	0,225 (0,012)	0,177 (0.012)	0,216 (0.009)
	Gordon	0,487 (0,023)	0,509 (0,013)	0,0463 (0,012)	0,508 (0,032)	0,372 (0,018)	0,479 (0,012)	0,359 (0,011)	0,458 (0,021)	**0,537** (0,017)	0,248 (0,013)	0,475 (0,011)
	Nutt2	0,010 (0,011)	**0,460** (0,019)	0,093 (0,021)	0,043 (0,023)	0,275 (0,012)	0,039 (0,021)	0,267 (0,014)	0,139 (0,021)	0,088 (0,003)	0,259 (0,008)	0,188 (0,009)
	Shipp	0,345 (0,023)	0,250 (0,012)	0,314 (0,013)	0,285 (0,018)	0,249 (0,011)	0,250 (0,17)	**0,672** (0,022)	0,371 (0,012)	0,470 (0,031)	0,345 (0,013)	0,331 (0,011)
cDNA	Alizadeh1	0,403 (0,011)	0,331 (0,019)	0,398 (0,022)	0,331 (0,021)	0,319 (0,012)	0,301 (0,020)	0,848 (0,009)	**0,973** (0,014)	0,045 (0,017)	0,649 (0,008)	0,780 (0,017)
	Chen	0,291 (0,031)	0,289 (0,021)	0,282 (0,023)	0,288 (0,018)	0,299 (0,022)	0,280 (0,017)	**0,793** (0,027)	0,786 (0,019)	0,255 (0,023)	0,288 (0,022)	0,100 (0,012)
	Khan	0,574 (0,021)	0,533 (0,018)	0,542 (0,023)	0,377 (0,022)	0,460 (0,017)	0,524 (0,013)	0,503 (0,023)	**0,710** (0,021)	0,421 (0,019)	0,455 (0,011)	0,495 (0,021)

Tabela 5.4: Média e desvio padrão (entre parênteses) do índice de FR (maiores valores em negrito) para os métodos de agrupamento e todos os conjuntos de dados.

A fim de fazer uma discussão sobre a relevância de variáveis/genes, foi selecionado um conjunto de dados de cada tecnologia. Esta seleção baseou-se no volume de informações disponíveis sobre os dados e o nível de dificuldade dos conjuntos. Portanto, para Affymetrix, escolhemos `Chowdary` e para cDNA escolhemos `Alizadeh1`.

`Chowdary` é um conjunto de dados que contém pares comparáveis de 104 snap congelados. Destes 104, 62 RNA foram conservados a partir de tecidos com tumores de mama (classe `B`) e outras 42 amostras foram de tumores de cólon do tipo Dukes'B (classe `C`). Todas as amostras são descritas por 182 genes (Chowdary et al., 2006). `Alizadeh1` consiste de amostras de expressões gênicas de linfomas doentes de adultos: *Diffuse Large B-Cell Lymphoma* (DLBCL). Tem duas formas distintas de DLBCL, contendo 21 amostras de `DLBCL1` e 21 amostras de `DLBCL2` deste conjunto de dados. Subgrupos de genes de DLBCL são distintos quando expresso durante a diferenciação de células B (`DLBCL1`) ou ativação (`DLBCL2`). Estas 42 amostras são descritas por 1095 genes (Alizadeh et al., 2000).

Para cada método multivariado e conjunto de dados, as médias do graus de pertinência multivariados a_{ij} foram calculadas usando o Algoritmo 12. A partir dessas médias, foram selecionados os cinco maiores valores para cada grupo e foram identificadas as variáveis/genes correspondentes — ver Tabela 5.5.

	Método	Grupo 1					Grupo 2				
		1	2	3	4	5	1	2	3	4	5
Chowdary	MFCM	0,0762 ZG16	0,0599 CRIPTO3	0,0569 DEFA6	0,0443 CCL20	0,0424 AOC1	0,0937 KRT14	0,0884 KRT15	0,0799 SCGB1D2	0,0770 NPY1R	0,0763 HLA-DQA1
	MFCM-G	0,0646 ZG16	0,0594 CRIPTO3	0,0557 DEFA6	0,0520 CCL20	0,0512 AOC1	0,1542 KRT14	0,1421 KRT15	0,1179 SCGB1D2	0,1057 NPY1R	0,1005 HLA-DQA1
	MFCM-D	0,0166 RPS4Y1	0,0151 CRIPTO3	0,0147 LEFTY1	0,0144 ZG16	0,0141 KRT6B	0,0212 KRT6B	0,0187 KRT14	0,0185 GABRP	0,0185 RPS4Y1	0,0184 PIP
	MPFCM	0,0867 ZG16	0,0709 CRIPTO3	0,0596 DEFA6	0,0487 CCL20	0,0462 AOC1	0,1128 KRT14	0,1079 KRT15	0,0988 SCGB1D2	0,0954 USH1C	0,0948 NPY1R
	MPFCM-D	0,2105 ZG16	0,1317 DEFA6	0,0833 NOX1	0,0798 CRIPTO3	0,0783 MEP1A	0,4026 KRT14	0,3808 GATA3	0,1065 SCGB1D2	0,0698 SCGB2A2	0,0172 GCNT3
Alizadeh1	MFCM	0,1723 CD83	0,1723 CSNK1G2	0,1723 WHSC1	0,1723 XKR6	0,1723 MORC2-AS1	0,1865 MEN1	0,0859 DNAH1	0,0859 WHSC1	0,0859 RNLS	0,0722 PRR3
	MFCM-G	0,2251 CD83	0,215 CSNK1G2	0,2113 WHSC1	0,2041 XKR6	0,2002 MORC2-AS1	0,8482 MEN1	0,0701 DNAH1	0,0489 WHSC1	0,0307 RNLS	0,0018 PRR3
	MFCM-D	0,1723 CD83	0,1723 CSNK1G2	0,1723 WHSC1	0,1723 XKR6	0,1723 MORC2-AS1	0,1865 MEN1	0,0859 DNAH1	0,0859 WHSC1	0,0859 RNLS	0,0722 PRR3
	MPFCM	0,2477 LOC401374	0,1434 CD83	0,1434 CSNK1G2	0,1434 WHSC1	0,1434 XKR6	0,8170 HLA-E	0,1234 TINF2	0,0396 SHMT2	0,0042 CLC	0,0031 CD40
	MPFCM-D	0,1162 SNX14	0,1118 C10ORF46	0,1094 TIAL1	0,1070 TM4SF2	0,0994 BNIP3	0,1264 CHD3	0,1237 TMEM41B	0,0807 GPR160	0,0713 CRHBP	0,0654 LBD29

Tabela 5.5: Valores selecionados de a_{ij} e correspondentes genes.

Inicialmente serão analisados os resultados a respeito do conjunto de dados `Chowdary` e depois de `Alizadeh1`. Para fazer a análise, o resultado do método multivariado que obteve o maior valor de FR foi selecionado, para cada conjunto de dados (MFCM em `Chowdary` e MFCM-G em `Alizadeh1`). No caso do `Chowdary`, quando consideramos genes no grupo 1, de acordo com estudos sobre estrógeno e células de câncer de mama, `ZG16` é reprimido por meio de sinalização do estrogénio em células de câncer da mama (Hsu et al., 2010). `CRIPTO3` é expresso na maioria dos tipos de câncer e está ausente ou apenas fracamente expresso em células normais. Este gene foi encontrado em tecidos da mama e câncer do pulmão (Sun et al., 2008). `DEFA6` pertence à família *Alpha Defensin*. *Alpha Defensins* em humanos aumentam a produção de citocinas e quimiocinas (a superfamília de pequenas proteínas do tipo citocina) e pode promover a metástase do câncer de mama (Muller et al., 2001). `CCL20` (pertence à família *Chemokine Ligands*) pode ser um fator envolvido na ontogenia [1] do carcinoma da mama (Marsigliante et al., 2013) uma vez que a metástase do câncer da mama são as fontes de *Ligands* mais abundantes (Muller et al., 2001). `AOC1` (também chamado DAO) é influenciada pelo ciclo menstrual, assim, as mulheres tem uma mais alta concentração de `DAO` do que os homens. Além disso, níveis de `DAO` estão intimamente associados com as mulheres saudáveis na pré-menopausa (Hamada et al., 2013). Os resultados obtidos estão, portanto, de acordo com a descrição da classe de tecidos de **câncer da mama** (`B`) de `Chowdary`.

Para o grupo 2, `KRT14` é expresso imunocitoquimicamente tanto em SP (*Side Population* — é uma sub-população de células que é distinta da população principal na base dos marcadores utilizados) e células não-SP e um estudo demonstrou que cânceres do sistema gastrointestinal

[1] Toda a sequência de eventos envolvidos no desenvolvimento de um organismo individual.

contêm células SP que mostram algumas características de células-tronco (Haraguchi et al., 2006). KRT15 pertence à mesma família de KRT14 e são reprimidos por Cdx2 (essencial para o desenvolvimento intestinal e desempenha um papel importante no câncer do cólon) (Liu et al., 2006). SCGB1D2 (ou Lipophilin B) é expresso no cólon (Zafrakas et al., 2006). O gene NPY1R pertence à família NPY onde níveis de NPY diferem entre os subgrupos de síndrome do intestino irritável (Simrén et al., 2003). HLA-DQA1 está fortemente associado com a doença celíaca, que é uma das mais frequentes doenças do intestino delgado (Rujner et al., 2001). Os resultados obtidos estão, portanto, de acordo com a descrição da classe de tecidos de **câncer do cólon** (C) de Chowdary. Em resumo, estes genes mostam ser relevantes para descrever os dois grupos de tecidos de Chowdary.

Passemos agora para a análise do conjunto de dados Alizadeh1. Quando se considera o grupo 1 (DLBCL1), CD83 é membro da família de genes CD, que pertence a uma grande família de genes chamada *Cluster of Differentiation* (ou Grupamento de Diferenciação). Genes CD permitem maturação e diferenciação de células em um subconjunto de DLBCL (Chang et al., 2007). CSNK1G2 é altamente expresso no Centro Germinativo (GC) de células normais do tipo B (Nishiu et al., 2004). XKR6 está relacionado com a diferenciação de células (Daniels et al., 2004; Redman et al., 1999). Portanto, estes genes estão em conformidade com a descrição da classe de tecidos de **diferenciação de células B** (DLBCL1).

Os genes relacionados com o grupo 2 (DLBCL2) têm um comportamento diferente. O gene com o maior valor de a_{ij} deste grupo é MEN1 que codifica *menin*. Estudos recentes revelaram que *menin* interage com uma variedade de proteínas, incluindo JunD (uma proteína de factor de transcrição) e que reprime a transcrição de JunD- e NF-kB-activado (Sowa et al., 2003), onde as ligações usadas por NF-kB é activado em muitas grandes células do linfoma, especialmente aqueles do tipo ABC (Gatter e Pezzella, 2010). RNLS codifica a proteína *renalase* cujas funções são a sobrevivência e fator de crescimento (Xu et al., 2005). Além disso, *renalase* pode ativar ligações de proteínas *quinase* de uma forma que proteja as células contra a apoptose (morte celular programa) (Wang et al., 2014). Por isso, estes genes estão em conformidade com a descrição da classe de tecidos de **ativação das células B** (DLBCL2). Para este conjunto de dados, de acordo com a literatura, não foi possível concluir se os genes WHSC1 e MORC2-AS1 estão, de fato, associados a DLBCL1, nem se os genes DNAH1, WHSC1 e PRR3 tem relação com o grupo DLBCL2.

Portanto, no contexto de conjunto de dados reais relativos a tecidos cancerígenos descritos por expressões gênicas e usando a estratégia proposta, foi possível mostrar que os genes selecionados eram relevantes para descrever cada grupo, em conformidade com a verdadeira partição (a estrutura real da classe na base de dados).

6
CONCLUSÕES

A melhor maneira que a gente tem de fazer possível amanhã alguma coisa que não é possível de ser feita hoje é fazendo hoje aquilo que hoje pode ser feito. Mas se eu não fizer hoje aquilo que hoje pode ser feito e tentar fazer hoje aquilo que hoje não pode ser feito, dificilmente eu faço amanhã aquilo que hoje não pude fazer.

—PAULO FREIRE

6.1 Conclusão

Diversos métodos de agrupamento usando a abordagem difusa presentes na literatura foram descritos neste trabalho. Estes métodos são baseados na teoria de conjuntos difusos que posteriormente foram aplicados a algoritmos de reconhecimento de padrões, do qual o principal representante é o *Fuzzy C-Means* (FCM). A abordagem difusa pode ser aplicada em diversas categorias de métodos tais como hierárquicos, particionais e de otimização. Dentre os particionais, versões foram propostas por pesquisadores objetivanto suprir alguma das deficiências do FCM, como o FCM com distância adaptativa que busca encontrar grupos de forma não-esférica, ou o FCM robusto que tem o intuito de obter melhores resultados que o FCM na presença de dados aberrantes. Além disso, existem versões onde há uma união entre a abordagem difusa e a possibilística gerando o *Possibilistic Fuzzy C-Means* (PFCM), onde a proposta dos autores é unir as vantagens do FCM com o *Possibilistic C-Means* (PCM).

Embora estes métodos de agrupamento difuso presentes na literatura obtêm bons resultados para muitos problemas, eles não são capazes de indicar o grau proximidade que um objeto tem para os grupos segundo uma dada variável. Isto é, o grau de pertinência agrega a

informação de todas as variáveis para poder obter um valor final. Este trabalho propôs uma nova forma de interpretar os dados a partir da análise de cada variável separadamente. Para isto, o grau de pertinência foi decomposto, onde a pertinência de um objeto para cada grupo pode ser representado por um vetor tão grande quanto o número de variáveis. Desta forma, o usuário tem à disposição mais informação para analisar os dados e o algoritmo que usa essa abordagem pode encontrar uma partição difusa com maior qualidade. Foram apresentados cinco métodos difusos multivariados, onde os três primeiros são baseados no FCM e os dois últimos são baseados no PFCM. Além disso, foram acrescentados índices de interpretação para avaliar a qualidade do agrupamento para cada grupo e variável a partir da partição difusa obtida por cada método proposto.

A fim de avaliar o desempenho dos algoritmos propostos, um estudo comparativo com seis métodos de agrupamento difuso foi realizado. O primeiro método é o Gustafson-Kessel que pode identificar grupos de diferentes formas. O segundo é o conhecido FCM, o qual é capaz de identificar grupos esféricos, uma vez que utiliza a distância Euclidiana. O terceiro método tem um parâmetro que fornece informações sobre a influência de variáveis individuais sobre os grupos detectados. O quarto é o *Possibilistic Fuzzy C-Means* que une tanto as características da abordagem possibilística quanto difusa. O quinto método é uma versão do FCM usando distâncias adaptativas que é capaz de identificar grupos de diferentes formas e tamanhos. O sexto é uma versão do FCM com uma ponderação por variável. A utilização do índice difuso surge da necessidade de avaliar partições difusas de uma forma mais apropriada. O uso do índice de Rand original (para partições rígidas) na comparação de partiçoes difusas poderia levar à perda de informação. Para fazer o estudo comparativo entre os métodos, experimentos foram realizados, os quais podem ser divididos em duas partes.

A primeira parte dos experimentos trata de configurações sintéticas. Cada uma contém classes arranjadas de acordo com uma distribuição normal bivariada. Ao total foram feitas treze configurações, divididas de acordo com os formatos das classes, presença de correlação ou dados ruidosos. Para avaliar quantitavamente os métodos, um experimento Monte Carlo foi realizado e os resultados mostraram que os métodos multivariados sem ponderação e com ponderação no grau de pertinência foram melhores para configurações com classes elípticas, enquanto que o método multivariado com ponderação na distância obteve melhores resultados onde predominava classes esféricas. Além disso, foi analisado o desempenho dos métodos com variáveis dependentes em diferentes situações. Para a abordagem possibilística, configurações com diferentes percentuais de dados ruidosos foram criadas. Em resumo, os resultados mostraram que os métodos multivariados são preferíveis quando as variáveis são independentes e possuem variabilidades intra-classe diferentes. A segunda parte dos experimentos trata de

conjuntos de dados obtidos no sítio da UCI, mostrando que os métodos propostos também são eficientes em dados reais. Por fim, foi apresentado a utilidade do índices de interpretação baseados nas respostas fornecidas pelos métodos propostos.

Também fez parte deste trabalho a análise dos graus difusos multivariados. Uma metodologia foi desenvolvida para interpretar esses graus. Inicialmente, foi apresentado um conjunto de dados sintéticos mostrando que os graus multivariados podem fornecer informações importantes a respeito da partição. Para fazer a análise, aplicações com dados biológicos a respeito de genes foram utilizadas, onde cada gene é uma variável. O primeiro conjunto Chowdary trata de tecidos com tumores de mama e de cólon. Os genes selecionados a partir dos graus multivariados apresentam ser relevantes para descrever os dois grupos de tecidos de Chowdary. O segundo conjunto de dados Alizadeh1 consiste de amostras de expressões gênicas de linfomas doentes de adultos. Para alguns dos genes selecionados para esse conjunto, não foi possível associá-los ao tipo de linfoma, mas por outro lado outros mostraram que são importantes para descrever tecidos com esse tipo de câncer. Portanto, no contexto de conjunto de dados reais relativos a tecidos cancerígenos descritos por expressões gênicas e usando a estratégia proposta, foi possível mostrar que os genes selecionados eram relevantes para descrever cada grupo, em conformidade com a verdadeira partição (a estrutura real da classe na base de dados).

6.2 Publicações

Os seguintes artigos foram publicados em revistas internacionais decorrentes dos métodos propostos neste trabalho:

1. Pimentel, B. A., & De Souza, R. M. C. R. (2013). A Multivariate Fuzzy C-Means Method. Applied Soft Computing, 13(4), 1592-1607.

2. Pimentel, B. A., & De Souza, R. M. C. R. (2014). A Weighted Multivariate Fuzzy C-Means Method in Interval-Valued Scientific Production Data. Expert Systems with Applications, 41(7), 3223-3236.

3. Pimentel, B. A., & De Souza, R. M. C. R. (2016). Multivariate Fuzzy C-Means algorithms with weighting. Neurocomputing, 174, 946-965.

O primeiro trata da introdução do método de agrupamento difuso onde os graus de pertinência são multivariados. Um estudo foi realizado com dados representados por pontos no espaço e foram feitos experimentos com dados sintéticos e conjuntos da UCI. O segundo artigo descreve a versão do método difuso multivariado com ponderação no grau de pertinência. Neste

artigo, foi feita uma aplicação com dados de produção científica brasileira representados por meio de dados simbólicos do tipo intervalo. O terceiro artigo engloba as três versões do método de agrupamento difuso multivariado baseado no FCM. Ele introduz os métodos com ponderação no grau de pertinência e na distância para dados representados por pontos no espaço e apresenta os índices de interpretação. São feitos experimentos com dados sintéticos e reais. Além disso, foi apresentada a utilização dos índices de interpretação para um conjundo de dados da UCI.

Em conferência internacional, foi publicado o seguinte artigo:

- Pimentel, B. A., Souto, M. C. P.; & De Souza, R. M. C. R. (2017). Interpreting Multivariate Membership Degrees of Fuzzy Clustering Methods: a Strategy. IEEE International Joint Conference on Neural Networks (IJCNN), .

Este artigo, diferentemente dos acima apresentados, propõe uma forma de interpretar os graus de pertinência multivariados a partir da matriz difusa final obtida pelo método multivariado. Neste artigo, a relevância de uma variável para um determinado grupo é medida e utilizada em aplicações com dados biológicos. Mostrou-se que graus multivariados podem identificar quais grupos de genes (variáveis) são mais relevantes para descrever um tecido cancerígeno (grupo).

Outro artigo publicado em revista internacional que não faz parte da tese, mas tem relação com o tema é o apresentado a seguir:

- Maciel, D. B. M., Amaral, G. J. A., De Souza, R. M. C. R, Pimentel, B. A. (2017). Multivariate Fuzzy K-Modes Algorithm. Pattern Analysis and Applications, 20(1), 59-71.

Também no contexto de método difuso, este artigo apresenta a abordagem multivariada no método *Fuzzy K-Modes*, criando assim um novo método multivariado. Os resultados mostraram que, maior a heterogeneidade (seja pelo número de categorias ou pela distribuição) entre as variáveis que descrevem os objetos, maior será a qualidade de agrupamento resultante da aplicação do método multivariado na base de dados.

Artigo publicado durante o doutorado em revista internacional relacionado ao contexto de métodos difusos:

- Silva Filho, T. M., Pimentel, B. A., De Souza, R. M. C. R., & Oliveira, A. L. I. (2015). Hybrid Methods for Fuzzy Clustering based on Fuzzy C-Means and Improved Particle Swarm Optimization. Expert Systems with Applications, 42(17), 6315-6328.

Este artigo propõe combinações do método *Fuzzy C-Means* e *Improved Particle Swarm Optimization* (versão do *Particle Swarm Optimization* com ajuste automático de parâmetros) para o problema de agrupamento difuso.

6.3 Trabalhos Futuros

Baseado no estudo realizado nesse trabalho, algumas observações importantes a respeito dos métodos propostos foram identificadas.

1. **Variáveis dependentes**: em muitos problemas do mundo real, as variáveis envolvidas na análise dos dados possuem alguma relação de dependência. Por exemplo, no contexto de Economia, normalmente o preço de um produto diminui a medida que sua demanda no mercado aumenta; na Medicina, o número de pacientes com algum tipo de problema cardiovascular aumenta a medida que o grau de sedentarismo, obesidade ou estresse de um determinado paciente também aumenta. Entretanto, os métodos multivariados supõem que as variáveis analisadas não possuem qualquer dependência entre si. Desta forma, surge a necessidade de propor extensões destes métodos de forma que sejam mais robustos a dados com variáveis dependentes;

2. **Dados não linearmente separáveis**: outra suposição que os métodos propostos neste trabalho fazem é que os grupos das partições podem ser lineramente separáveis. Isso acontece porque todos os métodos propostos herdam essa característica do FCM. Entretanto, assim como variáveis dependentes, podem existir problemas reais onde não é possível encontrar grupos através de retas. Desta forma, com o objetivo de deixar os métodos multivariados mais robustos a dados não linermante separáveis, é necessário que a metodologia aplicada ao FCM também seja utilizada em outros métodos que sejam comprovadamente melhor aplicados neste tipo de dados;

3. ***Multi-view***: uma representação bastante comum dos dados é através de uma única matriz, onde cada linha representa um objeto e as colunas representam variáveis que descrevem o objeto. Por outro lado, essa representação nem sempre é a mais adequada dependendo do tipo de problema, já que ela não considera que objetos podem ser descritos de diferentes formas. Por exemplo, na análise de dados sobre páginas da *web*, uma mesma página pode ser representada por diferentes conjuntos de palavras dependendo do contexto utilizado. Os métodos propostos fazem a suposição que os dados são representados de uma única perspectiva, isto é, por um único conjunto de variáveis. Desta forma, é interessante que o conceito de agrupamento *multi-view* seja aplicado aos métodos para se tornarem mais robustos a diferentes representações.

REFERÊNCIAS

Alberts, Bruce et al. (2002). Molecular Biology of the Cell; Fourth Edition.

Alizadeh, A. A. et al . (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403(6769), 503-511.

Altman R B, Building successful biological databases, Brief. (2004). Bioinformatics 5 (1): 4-5.

Anderson, D. T., Bezdek, J. C., Popescu, M., & Keller, J. M. (2010). Comparing fuzzy, probabilistic, and possibilistic partitions. Fuzzy Systems, IEEE Transactions on, 18(5), 906-918.

Anderson, J. R. (1986). Machine learning: An artificial intelligence approach (Vol. 2). R. S. Michalski, R. S. Michalski, & T. M. Mitchell (Eds.). Morgan Kaufmann.

Anzai, Y. (1992). Pattern Recognition & Machine Learning. Morgan Kaufmann.

Atla, Abhinav et al. Sensitivity of different machine learning algorithms to noise. Journal of Computing Sciences in Colleges, v. 26, n. 5, p. 96-103, 2011.

Attwood, T. K., Gisel, A., Bongcam-Rudloff, E., & Eriksson, N. E. (2011). Concepts, historical milestones and the central place of bioinformatics in modern biology: a European perspective. INTECH Open Access Publisher.

Babu, G. P., & Murty, M. N. (1994). Clustering with evolution strategies. Pattern recognition, 27(2), 321-329.

Bache, K., & Lichman, M. (2016). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Acessible: http://archive.ics.uci.edu/ml.

Bai, L., Liang, J., Sui, C., & Dang, C. (2013). Fast global k-means clustering based on local geometrical information. Information Sciences, 245, 168-180.

Baldi, P. & Brunak, S. (2001). Bioinformatics: the Machine Learning approach. MIT Press, segunda edição.

Bezdek, J. C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.

Bhaskar, H., Hoyle, D. C., & Singh, S. (2006). Machine learning in bioinformatics: A brief survey and recommendations for practitioners.Computers in biology and medicine, 36(10), 1104-1125.

Blum, A., & Chawla, S. (2001). Learning from labeled and unlabeled data using graph mincuts. In Proceedings of the Eighteenth International Conference on Machine Learning, 19-26.

Boeva, V. (2013). Clustering approaches for dealing with multiple DNA microarray datasets. Journal of Computational Science. 5(3): 368-376.

Brenner, S. et al. (2000). Gene expression analysis by Massive Parallel Signature Sequencing (MPSS) on microbead array. Nat. Biotechnol., 18(6):630–640.

Breve, F., Zhao, L., Quiles, M., Pedrycz, W., & Liu, J. (2012). Particle competition and cooperation in networks for semi-supervised learning. IEEE Transactions on Knowledge and Data Engineering, 24(9), 1686-1698.

Brouwer, R. K. (2009). Extending the rand, adjusted rand and jaccard indices to fuzzy partitions. Journal of Intelligent Information Systems, 32(3), 213-235.

Butler, J. M. (2005). Forensic DNA typing: biology, technology, and genetics of STR markers. Academic Press.

Campello, R. J. (2007). A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. Pattern Recognition Letters, 28(7), 833-841.

Chowdary, D. et al. Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. The journal of molecular diagnostics, vol. 8, n. 1, pp. 31-39, 2006.

Celeux, G., Diday, E., Govaert, G., Lechevallier, Y., & Ralambondrainy, H., (1989). Classification Automatique des Donnees. Bordas, Paris.

Chang, K.-C. et al. Distribution patterns of dendritic cells and T cells in diffuse large B-cell lymphomas correlate with prognoses. Clinical Cancer Research, vol. 13, n. 22, pp. 6666-6672, 2007.

Chatterjee, S., Banerjee, A., Acharya, T., & Maity, S. P. (2014). Fuzzy C-Means Clustering in Energy Detection for Cooperative Spectrum Sensing in Cognitive Radio System. In Multiple Access Communications (pp. 84-95). Springer International Publishing.

Chen, C. -W. The stability of an oceanic structure with T-S fuzzy models. Mathematics and Computers in Simulation, 80(2): 402-426, 2009.

Chen, C. Y., Lin, J. W., Lee, W. I., & Chen, C. -W. Fuzzy control for an oceanic structure: A case study in time-delay TLP system. Journal of Vibration and Control, 16(1): 147-160, 2010.

Chen, C. -W. Modeling and fuzzy PDC control and its application to an oscillatory TLP structure. Math. Problems in Engineering. An Open Access, 2010.

Chen, C. -W. Application of fuzzy-model-based control to nonlinear structural systems with time delay: an LMI method. Journal of Vibration and Control, 16, 1651-1672, 2010.

Choi, J. K., Yu, U., Kim, S., & Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. Bioinformatics, 19(suppl 1), i84-i90.

Cox, E. (2005). Fuzzy modeling and genetic algorithms for data mining and exploration. Academic Press. Chicago.

Daniels, G. L. et al. Blood group terminology 2004: from the International Society of Blood Transfusion committee on terminology for red cell surface antigens. Vox sanguinis, vol. 87, n. 4, pp. 304-316, 2004.

De Carvalho, F. A. T., De Souza, R. M., Chavent, M., & Lechevallier, Y. (2006). Adaptive Hausdorff distances and dynamic clustering of symbolic interval data. Pattern Recognition Letters, 27(3), 167-179.

De Carvalho, F. A. T., Tenorio, C. P., & Cavalcanti, N. L. J. (2006). Partitional fuzzy clustering methods based on adaptive quadratic distances. Fuzzy sets and systems, 157(21), 2833-2857.

De Carvalho, F. A. T., De Souza, R. M. C. R., Chavent, M., & Lechevallier, Y. (2006). Adaptive Hausdorff distances and dynamic clustering of symbolic interval data. Pattern Recognition Letters, 27(3), 167-179.

De Carvalho, F. A. T., Brito, P., & Bock, H. H. (2006). Dynamic clustering for interval data based on L2 distance. Computational Statistics, 21(2), 231-250.

De Souza, R. M. C. R.; De Carvalho, F. de A. T. Clustering of interval data based on city–block distances. Pattern Recognition Letters, v. 25, n. 3, p. 353-365, 2004.

De Souto, M. C. P., Lorena, A. C., Delbem, A. C. B., & de Carvalho, A. C. P. L. F. (2003). Técnicas de aprendizado de máquina para problemas de biologia molecular. Sociedade Brasileira de Computaçao.

De Souto, M. C. P. et al. Clustering cancer gene expression data: a comparative study. BMC bioinformatics, vol. 9, n. 1, pp. 1, 2008.

Dembélé D (2013). A flexible microarray data simulation model. Microarrays, 2(2):115-130.

Diday, E., & Noirhomme-Fraiture, M. (2008). Symbolic Data Analysis And The Sodas Software. Wiley-Interscience Publishers.

Diday, E. & Simon, J. J. (1976). Clustering Analisys. K. S. Fu (ed), Digital Pattern Recognition, 47-94.

Diday, E. & Gouvaert, G. Classification Automatique avec Distances Adaptatives. R. A. I. R. O. Informatique Computer Science, 11(4), 329-349.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern Classification, John Wiely and Sons, New York, (Chapter 1).

Duggan, D. J. et al. (1999). Expression profiling using cDNA microarrays. Nature Genetics, 21, 10-14.

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, Journal of Cybernetics 3, 32-57.

D'Urso, P., & Giordani, P. (2006). A Robust Fuzzy K-Means Clustering Model for Interval Valued Data. Computational Statistics, 21, 251-269.

Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). Cluster Analysis. Wiley Series in Probability and Statistics.

Ferreira, M. R., & De Carvalho, F. D. A. (2014). Kernel Fuzzy C-Means With Automatic Variable Weighting. Fuzzy Sets and Systems, 237, 1-46.

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. The computer journal, 41(8), 578-588.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the american statistical association, 32(200), 675-701.

Frigui, H., Hwang, C., & Chung-Hoon Rhee, F. (2007). Clustering and aggregation of relational data with applications to image database categorization. Pattern Recognition, 40(11), 3053-3068.

Gan, G., Ma, C., & Wu, J. (2007). Data clustering: theory, algorithms, and applications (Vol. 20). Siam.

Gatter, K.; Pezzella, F. Diffuse large B-cell lymphoma. Diagnostic Histopathology, vol. 16, n. 2, pp. 69-81, 2010.

Golub, T. R. et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science,286(5439), 531-537.

Groenen, P. J. F., e K. Jajuga (2001). Fuzzy clustering with squared Minkowsky distances. Fuzzy Sets and Systems 120, 227-237.

Gustafson, D.E. e Kessel, W.C. Fuzzy clustering with a fuzzy covariance matrix. In Proc. IEEE CDC, pages 761–766, San Diego, CA, USA, 1979.

Hamada, Y. et al. Effect of the menstrual cycle on serum diamine oxidase levels in healthy women. Clinical biochemistry, vol. 46, n. 1, pp. 99-102, 2013.

Haraguchi, N. et al. Cancer stem cells in human gastrointestinal cancers. Human cell, vol. 19, n. 1, pp. 24-29, 2006.

Hathaway, R. J., Bezdek, J. C., & Hu, Y. (2000). Generalized Fuzzy c-Means Clustering Strategies Using Lp Norm Distance. IEEE Transactions on Fuzzy Systems, 8(5), 576-572.

Hruschka, E. R., Campello, R. J. G. B., Freitas, A. A., & De Carvalho, A. P. L. F. (2009). A survey of evolutionary algorithms for clustering. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 39(2), 133-155.

Hsu, P.-Y. et al. Estrogen-mediated epigenetic repression of large chromosomal regions through DNA looping. Genome research, vol. 20, n. 6, pp. 733-744, 2010.

Hu, Qinghua, Shuang An, & Daren Yu. Soft fuzzy rough sets for robust feature evaluation and selection. Information Sciences 180.22: 4384-4400, 2010.

Huang, J. Z., Ng, M. K., Rong, H., & Li, Z. (2005). Automated variable weighting in k-means type clustering. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(5), 657-668.

Hullermeier, E., Rifqi, M., Henzgen, S., & Senge, R. (2012). Comparing fuzzy partitions: A generalization of the Rand index and related measures. IEEE Transactions on Fuzzy Systems, 20(3), 546-556.

Ichikawa, K., & Morishita, S. (2014). A simple but powerful heuristic method for accelerating k-means clustering of large-scale data in life science.IEEE/ACM Transactions on Computational Biology and Bioinformatics, 99, 1-12.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323.

Jain, A. K. & Dubes, R. C. (1988). Algorithms for Clustering Data. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651-666.

Jajuga, K. (2001). L1-norm based fuzzy clustering. Fuzzy Sets and Systems, 39, 63-83.

Ji, Zexuan et al. Interval-valued possibilistic fuzzy C-means clustering algorithm. Fuzzy Sets and Systems, v. 253, p. 138-156, 2014.

Jiang, D., Tang, C., & Zhang, A. (2004). Cluster analysis for gene expression data: a survey. Knowledge and Data Engineering, IEEE Transactions on, 16(11), 1370-1386.

Karsch-Mizrachi, I., Nakamura, Y., & Cochrane, G. (2012). The international nucleotide sequence database collaboration. Nucleic acids research, 40(D1), D33-D37.

Keller, A., & Klawonn, F. (2000). Fuzzy clustering with weighting of data variables. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 8(6), 735-746.

Keller, J., Krisnapuram, R., & Pal, N. R. (2005). Fuzzy models and algorithms for pattern recognition and image processing (Vol. 4). Springer.

Khan, J. et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature medicine, 7(6), 673-679.

Krishnamurthy, A., Balakrishnan, S., Xu, M., & Singh, A. (2012). Efficient active algorithms for hierarchical clustering. Proccedings of the 29th Internacional Conference on Machine Learning. arXiv preprint arXiv:1206.4672.

Krishnapuram, R., & Keller, J. M. (1993). A possibilistic approach to clustering. IEEE Transactions on Fuzzy Systems, 1(2), 98-110.

Kuo, R. J., Chen, S. S., Cheng, W. C., & Tsai, C. Y. (2013). Integration of artificial immune network and K-means for cluster analysis. Knowledge and Information Systems, 1-17.

Leski, Jacek M. Fuzzy c-ordered-means clustering. Fuzzy Sets and Systems, v. 286, p. 114-133, 2016.

Liu, T. et al. Regulation of Cdx2 expression by promoter methylation, and effects of Cdx2 transfection on morphology and gene expression of human esophageal epithelial cells. Carcinogenesis, vol. 28, n. 2, pp. 488-496, 2006.

Lu, C., Xiao, S., & Gu, X. (2014). Improving fuzzy C-means clustering algorithm based on a density-induced distance measure. The Journal of Engineering, 1(1).

Maciel, D. B. M., Amaral, G. J. A., De Souza, R. M. C. R, Pimentel, B. A. (2017). Multivariate Fuzzy K-Modes Algorithm. Pattern Analysis and Applications, 20(1), 59-71.

MacKay, D. J. (2003). Information theory, inference and learning algorithms. Cambridge university press.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability 1(14): 281-297.

Mann, H. B. & Whitney, D. R. (1947). On a test whether one of two random variables is stochastically larger than the other. Ann. Math. Statist., 18: 50-60.

Marsigliante, S.; Vetrugno, C.; Muscella, A. CCL20 induces migration and proliferation on breast epithelial cells. Journal of cellular physiology, vol. 228, n. 9, pp. 1873-1883, 2013.

Mitchell, T. (1997). Machine Learning. McGraw Hill, New York.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50(2), 159-179.

Muller, A. et al. Involvement of chemokine receptors in breast cancer metastasis. Nature, vol. 410, n. 6824, pp. 50-56, 2001.

Nazari, M., Shanbehzadeh, J., & Sarrafzadeh, H. (2013). Fuzzy c-means based on automated variable feature weighting. In Proceedings of the International MultiConference of Engineers and Computer Scientists, 1-5.

Nishiu, M. et al. Distinct pattern of gene expression in pyothorax-associated lymphoma (PAL), a lymphoma developing in long-standing inflammation. Cancer science, vol. 95, n. 10, pp. 828-834, 2004.

Pal, N. R., Pal, K., Keller, J. M., & Bezdek, J. C. (2005). A possibilistic fuzzy c-means clustering algorithm. IEEE Transactions on Fuzzy Systems, 13(4), 517-530.

Pal, N. R., Pal, K., & Bezdek, J. C. (1997). A mixed c-means clustering model. In Fuzzy Systems, Proceedings of the Sixth IEEE International Conference on (Vol. 1, pp. 11-21).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12, 2825-2830.

Petry, F., & Robinson, V. B. (2005). Fuzzy modeling with spatial information for geographic problems. Berlin etc.: Springer. Chicago.

Pimentel, B. A., & Souza, R. M. C. R. (2013). A Multivariate Fuzzy C-Means Method. Applied Soft Computing, 13(4), 1592-1607.

Pimentel, B. A., & Souza, R. M. C. R. (2014). A Weighted Multivariate Fuzzy C-Means Method in Interval-Valued Scientific Production Data. Expert Systems with Applications, 41(7), 3223-3236.

Pimentel, B. A., & de Souza, R. M. (2016). Multivariate Fuzzy C-Means algorithms with weighting. Neurocomputing, 174, 946-965.

Pimentel, B. A., Souto, M. C. P., & De Souza, R. M. C. R. (2017). Interpreting Multivariate Membership Degrees of Fuzzy Clustering Methods: a Strategy. IEEE International Joint Conference on Neural Networks (IJCNN), .

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66(336), 846-850.

Redman, C. M.; Russo, D.; Lee, S. Kell, Kx and the McLeod syndrome. Best Practice & Research Clinical Haematology, vol. 12, n. 4, pp. 621-635, 1999.

Rujner, J. et al. Coeliac disease and HLA-DQ 2 (DQA1* 0501 and DQB1* 0201) in patients with Turner syndrome. Journal of pediatric gastroenterology and nutrition, vol. 32, n. 1, pp. 114-115, 2001.

Ruspini, E. H. (1969). A new approach to clustering. Information and Control, 15, 22-32.

Setúbal, J. C. & Meidanis, J. (1997). Introduction to Computational Molecular Biology. PWS Publishing Company.

Silva Filho, T. M., Pimentel, B. A., De Souza, R. M. C. R., & Oliveira, A. L. I. (2015). Hybrid Methods for Fuzzy Clustering based on Fuzzy C-Means and Improved Particle Swarm Optimization. Expert Systems with Applications, 42(17), 6315-6328.

Simrén, M. et al. Abnormal levels of neuropeptide Y and peptide YY in the colon in irritable bowel syndrome. European journal of gastroenterology & hepatology, vol. 15, n. 1, pp. 55-62, 2003.

Slack, J. M. W (2014). Genes-A Very Short Introduction. Oxford University Press.

Sowa, H. et al. Inactivation of menin, the product of the multiple endocrine neoplasia type 1 gene, inhibits the commitment of multipotential mesenchymal stem cells into the osteoblast lineage. Journal of Biological Chemistry, vol. 278, n. 23, pp. 21058-21069, 2003.

Sun, C. et al. CRIPTO3, a presumed pseudogene, is expressed in cancer. Biochemical and biophysical research communications, vol. 377, n. 1, pp. 215-220, 2008.

Szekely, G. J., & Rizzo, M. L. (2005). Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. Journal of classification, 22(2), 151-183.

Torra, V. (2005). Fuzzy c-means for Fuzzy Hierarchical Clustering. In FUZZ-IEEE (pp. 646-651).

Verma, P., & Yadava, R. D. S. (2014). Application of Fuzzy c-Means Clustering for Polymer Data Mining for Making SAW Electronic Nose. In Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA). Springer International Publishing.

Wang, L. et al. Renalase prevents AKI independent of amine oxidase activity. Journal of the American Society of Nephrology, vol. 25, n. 6, pp. 1226-1235, 2014.

Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques. Elsevier.

Wu, K.-L. (2012). Analysis of parameter selections for fuzzy c-means. Pattern Recognition, 45, 407-415.

Xiong, Hui et al. Enhancing data analysis with noise removal. Knowledge and Data Engineering, IEEE Transactions on, v. 18, n. 3, p. 304-319, 2006.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on Neural Networks, 16(3), 645-678.

Xu, J. et al. Renalase is a novel, soluble monoamine oxidase that regulates cardiac function and blood pressure. The Journal of clinical investigation, vol. 115, n. 5, pp. 1275-1280, 2005.

Yang, M. S. (1993). A survey of fuzzy clustering. Mathematical and Computer modelling, 18(11), 1-16.

Yeh K., & Chen C. W. Stability analysis of interconnected fuzzy systems using the fuzzy Lyapunov method. Mathematical Problems in Engineering. An Open Access Journal, 2010.

Yu, J., Cheng, Q. & Huang, H. (2004). Analysis of the weighting exponent in the FCM, IEEE Transactions on Systems, Man, and Cybernetics - Part B, 34, 634-639.

Zadeh, L. A. (1965). Fuzzy sets. Information and control, 8(3), 338-353.

Zafrakas, M. et al. Expression analysis of mammaglobin A (SCGB2A2) and lipophilin B (SCGB1D2) in more than 300 human tumors and matching normal tissues reveals their co-expression in gynecologic malignancies. BMC cancer, vol. 6, n. 1, pp. 88, 2006.

Zhang, D. Q., & Chen, S. C. (2003). Kernel-based fuzzy and possibilistic c-means clustering. In Proceedings of the International Conference Artificial Neural Network (pp. 122-125).

GLOSSÁRIO

Agrupamento é uma técnica computacional de aprendizagem não-supervisionada cujo objetivo consiste em separar objetos em grupos, levando-se em conta as características que estes objetos possuem. É sinônimo de *Clustering*.

Algoritmo é o conjunto finito de instruções bem definidas e não ambíguas cujo principal objetivo é realizar uma tarefa ou resolver um problema pré-determinado.

Análise de Agrupamento é o nome dado ao conjunto de técnicas elaboradas para a formação de grupo de objetos sem levar em conta as classes aos quais esses objetos pertencem. Consiste em uma ferramenta para análise de dados onde são utilizadas estatísticas e heurísticas para descrever esses dados com o objetivo de obter alguma informação.

Bioinformática é o ramo da ciência que consiste em fazer uso de técnicas de informática aplicadas à análise de dados advindos de estudos relacionados à Biologia. É um campo interdiciplinar que envolve conhecimentos de Química, Física, Biologia, Ciência da Computação e Estatística para processar dados biológicos ou biomédicos.

Classe é um conjunto de objetos com rótulos. Esse conjunto é definido *a priori* fazendo parte da informação da configuração de dados ou pode ser o resultado da execução de um método de classificação (aprendizagem supervisionada).

Cluster é um conjunto de objetos não rotulados resultante de um método de agrupamento (não supervisionado) de forma que esses objetos possuem entre si um mais alto grau de similaridade quando comparados com outros objetos de um diferente *cluster*. É sinônimo de grupo.

Complexidade de tempo é a quantificação da porção de tempo tomada por um algoritmo para executar uma determinada tarefa em função do tamanho da sua entrada e é estimada pela contagem do número de operações elementares realizadas pelo algoritmo.

Complexidade de espaço é a estimativa da porção de espaço requerida por um algoritmo para resolver um problema onde o cálculo dessa estimativa leva em conta o tamanho dos dados fornecidos e quanto de espaço foi alocado para poder realizar as instruções do algoritmo.

Convergência é a disposição de um algoritmo quando a sequência de erros tende a zero, onde a sequência de erros consiste na diferença entre cada termo e o limite de uma sequência de valores. Nos algoritmos de otimização, cada termo é um valor do critério relativo a uma determinada iteração e o limite da sequência é a iteração atual do algoritmo.

Dados aberrantes são valores atípicos que se distanciam de outras observações resultantes de erro humano, de erro do instrumento, de mudanças no comportamento do sistema ou de características naturais dos dados por possuírem distribuições de cauda pesada. Podem ser *outliers* ou ruídos.

Difuso é a característica daquilo que pode ser representado de uma maneira mais abrangente. Que não apresenta limite preciso e é normalmente quantificado por valores contínuos.

Dissimilaridade é a diferença entre dois objetos. Quanto mais diferente forem, maior a dissimilaridade. Normalmente utiliza-se o conceito de distância para quantificar a dissimilaridade.

Função objetivo é a função de um ou mais elementos utilizada em métodos de otimização com o intuito de minimizá-la ou maximizá-la. Essa função mapeia um conjunto de entrada em valores reais. Desta forma, os métodos de otimização buscam modelar os problemas do mundo real e teórico através de função objetivo e encontrar soluções de forma que o valor fornecido por essa função seja o menor/maior possível.

Gene é uma região do ADN (Ácido Desoxirribonucleico) que contém código para a produção dos aminoácidos da cadeia polipeptídica (constituinte fundamental das proteínas) e as sequências reguladoras para a expressão. É a unidade molecular da hereditariedade formado por uma sequência específica de ácidos nucléicos.

Grau de pertinência é a medida referente à proximidade de um objeto a um protótipo de um determinado grupo segundo a abordagem difusa no contexto de agrupamento. É calculado sob a restrição da soma: os graus de pertinência de um dado objeto para todos os protótipo deve-se obter a soma igual a 1.

Grau de possibilidade é a medida referente à proximidade de um objeto a um protótipo de um determinado grupo segundo a abordagem possibilística no contexto de agrupamento. Neste tipo

de grau, a restrição da soma não é utilizada, de forma que soma dos graus de possibilidade de um dado objeto para todos os protótipos não é necessariamente igual a 1.

Grupo é a união de um ou mais objetos não rotulados. É constituído de tal forma que esses objetos possuem entre si um mais alto grau de similaridade quando comparados com outros objetos de um diferente grupo. É sinônimo de *cluster*.

Índice é um número obtido por meio de uma média, ou outro procedimento similar, capaz de representar um conjunto de valores. Portanto, índices de avalição ou de interpretação são números obtidos a partir de outros valores utilizados para avaliar ou interpretar, respectivamente, o elemento desejado.

Inter-grupo é a relação existente entre grupos distintos.

Intra-grupo é a relação existente entre elementos do mesmo grupo.

Iteração é o ato de repetir um processo com o objetivo de se atingir uma meta ou resultado desejado de forma os resultados obtidos em uma determinada iteração são usados na iteração imediatamente seguinte.

Objeto é cada entidade do conjunto de dados descrito por um conjunto pré-determinado de atributos (ou variáveis) as quais contem informações sobre medidas de características dos objetos. É o elemento a ser classificado ou agrupado por um método de aprendizagem de máquina.

Outlier é um valor atípico que se distancia de outras observações e que pertence a uma determinada classe.

Método é a constituição de uma série de passos codificados que se têm de tomar, de forma esquemática para atingir um determinado objetivo científico.

Multivariado é o que oferece múltipla variações. Em estatística, a Análise Multivariada trata de um conjunto de métodos que permite a análise simultânea de medidas múltiplas para cada objeto em análise, ajudando o pesquisador a tomar decisões melhores no contexto em que é dada a informação disponível para o conjunto de dados analisados.

Partição é o conjunto de dois ou mais grupos de forma que não exista interseção entre quaisquer pares de grupos. É o resultado da execução de um método de particionamento onde uma certa quantidade de objetos são alocados para grupos cuja quantidade é pré-definida de forma que cada objeto só pode ser alocado a um único grupo.

Peso é um valor com o qual se multiplicam certas grandezas, com a finalidade de lhes dar maior ou menor grandeza. Normalmente os pesos representam a quantificação da relevância de alguma entidade (por exemplo, peso para uma variável ou distância).

Possibilístico é a propriedade daquilo que pode ser interpretado usando possibilidades. Diferentemente do probabilístico, o possibilístico leva em conta a possibilidade de um evento acontecer, e não a probabilidade. Dessa forma, não considera que a soma das possibilidades de um evento seja necessariamente igual 1.

Ponderação é o ato de acrescentar pesos e, consequentemente, tornar possível que a entidade ponderada leve em conta diferentes medidas as quais informam a relevância a respeito de algo.

Protótipo é o representante de um determinado grupo. É descrito por um vetor o qual pode ser interpretado como um elemento central de uma coleção de objetos. Em alguns métodos como o *K-Means*, o protótipo é a média dos vetores de objetos sendo assim denominado centróide e é um elemento sintético. Por outro lado, no método *K-Medoids*, passa a ser chamado medóide e é um elemento do conjunto de dados.

Ruído é um valor atípico resultante de erro humano, de erro do instrumento ou de mudanças no comportamento do sistema sem estar associado a uma classe comprometendo o desempenho de métodos.

Sequência é uma função real cujo domínio é um conjunto contável totalmente ordenado (números natuais, por exemplo) e o contradomínio é considerado o conjunto dos números reais de forma que, para cada número natual, existe um número real resultante da aplicação da função ao primeiro.

Variável é uma característica ou aspecto de um objeto utilizado para descrevê-lo. Normalmente os objetos são descritos por várias variáveis onde cada variável possui um valor associado.

APÊNDICE A - PROVAS

Este parte destina-se a provas de expressões anunciadas no texto.

A.1 Prova da Proposição 3.2.1

Fixando o grau de pertinência u_{ijk}, o prótótipo y_{ij} que minimiza o critério J^1 é atualizado usando a seguinte expressão:

$$y_{ij} = \frac{\sum_{k=1}^{n} (u_{ijk})^m x_{kj}}{\sum_{k=1}^{n} (u_{ijk})^m}. \tag{A.1}$$

Prova. O critério J^1 sendo aditivo, o problema torna-se minimizar J_{ij}^1 como a seguir. Seja $d_{ijk} = (x_{kj} - y_{ij})^2$ ser a diatância entre o objeto $\mathbf{x}_k$ e o protótipo $\mathbf{y}_i$ para a variável j:

$$J_{ij}^1 = \sum_{k=1}^{n} (u_{ijk})^m (x_{kj} - y_{ij})^2. \tag{A.2}$$

J_{ij}^1 torna-se estacionário quando:

$$\frac{\partial J_{ij}^1}{\partial y_{ij}} = -2 \sum_{k=1}^{n} (u_{ijk})^m (x_{kj} - y_{ij}) = 0 \tag{A.3}$$

Isto é:

$$\sum_{k=1}^{n} (u_{ijk})^m x_{kj} - y_{ij} \sum_{k=1}^{n} (u_{ijk})^m = 0 \tag{A.4}$$

A partir disto, pode-se concluir que:

$$y_{ij} = \frac{\sum_{k=1}^{n} (u_{ijk})^m x_{kj}}{\sum_{k=1}^{n} (u_{ijk})^m}. \square \tag{A.5}$$

A.2 Prova da Proposição 3.2.2

Fixando o protótipo y_{ij}, o grau de pertinência u_{ijk} que minimiza o critério J^1 é atualizado usando a seguinte equação sob a restrição $\sum_{i=1}^{c}\sum_{j=1}^{p} u_{ijk} = 1$:

$$u_{ijk} = \left[\sum_{a=1}^{c}\sum_{b=1}^{p}\left(\frac{d_{ijk}}{d_{abk}}\right)^{\frac{1}{m-1}}\right]^{-1}. \tag{A.6}$$

Prova. O critério J^1 sendo aditivo, o problema torna-se encontrar, para algum objeto $\mathbf{x}_k$, o grau de pertinência u_{ijk} que minimiza

$$J_k^1 = \sum_{i=1}^{c}\sum_{j=1}^{p}(u_{ijk})^m d_{ijk}. \tag{A.7}$$

O método dos multiplicadores de Lagrange pode ser aplicado para encontrar um ótimo valor para J_k^1:

$$F_k(\lambda) = \sum_{i=1}^{c}\sum_{j=1}^{p}(u_{ijk})^m d_{ijk} - \mu\left(\sum_{i=1}^{c}\sum_{j=1}^{p}u_{ijk} - 1\right). \tag{A.8}$$

Assim, F_k é estacionário nas seguintes situações:

1. $\frac{\partial F_k}{\partial \mu} = \sum_{i=1}^{c}\sum_{j=1}^{p} u_{ijk} - 1 = 0$;
2. $\frac{\partial F_k}{\partial u_{ijk}} = m(u_{ijk})^{m-1} d_{ijk} - \mu = 0$.

Do segundo item, é possível calcular o valor do grau de pertinência u_{ijk} como a seguir:

$$u_{ijk} = \left(\frac{\mu}{m}\right)^{\frac{1}{m-1}}\left(\frac{1}{d_{ijk}}\right)^{\frac{1}{m-1}}. \tag{A.9}$$

De acordo com a restrição $\sum_{i=1}^{c}\sum_{j=1}^{p} u_{ijk} = 1$ e a Equação A.9, tem-se:

$$\sum_{a=1}^{c}\sum_{b=1}^{p} u_{abk} = \left(\frac{\mu}{m}\right)^{\frac{1}{m-1}}\sum_{a=1}^{c}\sum_{b=1}^{p}\left(\frac{1}{d_{abk}}\right)^{\frac{1}{m-1}} = 1. \tag{A.10}$$

Desta forma, obtem-se a seguinte expressão:

$$\left(\frac{\mu}{m}\right)^{\frac{1}{m-1}} = \frac{1}{\sum_{a=1}^{c}\sum_{b=1}^{p}\left(\frac{1}{d_{abk}}\right)^{\frac{1}{m-1}}}. \tag{A.11}$$

Substituindo a Equação A.11 na Equação A.9, tem-se:

$$u_{ijk} = \frac{1}{\sum_{a=1}^{c}\sum_{b=1}^{p}\left(\frac{1}{d_{abk}}\right)^{\frac{1}{m-1}}}\left(\frac{1}{d_{ijk}}\right)^{\frac{1}{m-1}}. \tag{A.12}$$

Finalmente, obtem-se:

$$u_{ijk} = \left[\sum_{a=1}^{c}\sum_{b=1}^{p}\left(\frac{d_{ijk}}{d_{abk}}\right)^{\frac{1}{m-1}}\right]^{-1}.\square \tag{A.13}$$

A.3 Prova da Proposição 3.2.3

A série $\omega_t = J^1(v_t)$ decresce a cada iteração e converge.

Prova. Segundo o algoritmo, as seguintes desigualdades (I) e (II) acontecem a cada iteração:

$$J^1(\mathbf{Y}^t,\mathbf{U}^t)\overbrace{\geq}^{(I)}J^1(\mathbf{Y}^{t+1},\mathbf{U}^t)\overbrace{\geq}^{(II)}J^1(\mathbf{Y}^{t+1},\mathbf{U}^{t+1}). \tag{A.14}$$

Considere $d(x_{kj},y_{ij}^{(t)}) = (x_{kj}-y_{ij}^{(t)})^2$ a distância entre o objeto x_{kj} e o protótipo y_{ij} na iteração t. A desigualdade (I) acontece porque

$$J^1(\mathbf{Y}^t,\mathbf{U}^t) = \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk}^{(t)})^m d(x_{kj},y_{ij}^{(t)}),$$

$$J^1(\mathbf{Y}^{t+1},\mathbf{U}^t) = \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk}^{(t)})^m d(x_{kj},y_{ij}^{(t+1)}),$$

e de acordo com a Proposição 3.2.1,

$$y_{ij}^{(t+1)} = \underbrace{argmin}_{y\in\Re}\sum_{k=1}^{n}(u_{ijk}^{(t)})^m d(x_{kj},y).$$

Além disso, a desigualdade (II) também acontece porque

$$J^1(\mathbf{Y}^{t+1},\mathbf{U}^{t+1}) = \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk}^{(t+1)})^m d(x_{kj},y_{ij}^{(t+1)}),$$

e de acordo com a Proposição 3.2.2,

$$u_{ijk}^{(t+1)} = \underbrace{argmin}_{u\in[0,1]} \sum_{k=1}^{n} (u)^m d(x_{kj}, y_{ij}^{(t+1)}).$$

Finalmente, uma vez que a série ω_t decresce e é limitado por $(J^1(v_t) \geq 0)$, ela converge. □

A.4 Prova da Proposição 3.2.4

A série $v_t = (\mathbf{Y}^t, \mathbf{U}^t)$ converge.

Prova. Considere que o ponto estacionário da série v_t é atingido na iteração $t = T$. Assim, pode-se dizer que $v_T = v_{T+1}$ e então $J^1(v_T) = J^1(v_{T+1})$, isto é, $J^1(\mathbf{Y}^T, \mathbf{U}^T) = J^1(\mathbf{Y}^{T+1}, \mathbf{U}^{T+1})$. A partir desta igualdade, a Proposição 3.2.3 pode ser reescrita como as igualdades (I) e (II):

$$J^1(\mathbf{Y}^T, \mathbf{U}^T) \overset{(I)}{=} J^1(\mathbf{Y}^{T+1}, \mathbf{U}^T) \overset{(II)}{=} J^1(\mathbf{Y}^{T+1}, \mathbf{U}^{T+1}) \tag{A.15}$$

Segundo a igualdade (I), $\mathbf{Y}^T = \mathbf{Y}^{T+1}$ porque $\mathbf{Y}$ é único minimizando J^1 quando $\mathbf{U}^T$ é fixo. Da segunda igualdade (II), $\mathbf{U}^T = \mathbf{U}^{T+1}$ porque $\mathbf{U}$ é único minimizando J^1 quando $\mathbf{Y}^{T+1}$ é fixo.

Finalmente, conclui-se que $v_T = v_{T+1}$. Esta conclusão acontece para todo $t \geq T$ e $v_t = v_T, \forall t \geq T$ e segue-se que v_t converge. □

A.5 Prova da Proposição 3.2.7

Fixado o grau de pertinência u_{ijk} e o protótipo y_{ij}, o peso γ_{ijk} que minimiza o critério J^2 é atualizado usando a seguinte equação sob a restrição $\sum_{i=1}^{c}\sum_{j=1}^{p} \gamma_{ijk} u_{ijk} = 1$:

$$\gamma_{ijk} = \frac{\left[(u_{ijk})^m d_{ijk}\right]^{\frac{1}{m+1}}}{\sum_{a=1}^{c}\sum_{b=1}^{p}\left[(u_{abk})^m d_{abk}\right]^{\frac{1}{m+1}} u_{abk}}. \tag{A.16}$$

Prova. Sabendo que o critério J^2 é aditivo e aplicando o método dos multiplicadores de Lagrange para encontrar um valor ótimo para J_k^2, tem-se:

$$F_k(\lambda) = \sum_{i=1}^{c}\sum_{j=1}^{p} (u_{ijk})^m (\gamma_{ijk})^{-m} d_{ijk} - \mu\left(\sum_{i=1}^{c}\sum_{j=1}^{p} \gamma_{ijk} u_{ijk} - 1\right) \tag{A.17}$$

F_k é estacionário nas seguintes situações:

1. $\frac{\partial F_k}{\partial \mu} = \sum_{i=1}^{c}\sum_{j=1}^{p} \gamma_{ijk} u_{ijk} - 1 = 0$;

2. $\frac{\partial F_k}{\partial \gamma_{ijk}} = -m(\gamma_{ijk})^{-m-1}(u_{ijk})^m d_{ijk} - \mu = 0.$

De acordo com o item 2, é possível reformular o peso do objeto x_k do grupo C_i para a variável j usando a seguinte equação:

$$\gamma_{ijk} = \left(-\frac{\mu}{m}\right)^{-\frac{1}{m+1}} \left(\frac{1}{(u_{ijk})^m d_{ijk}}\right)^{-\frac{1}{m+1}} \tag{A.18}$$

Uma vez que o peso está sob a restrição $\sum_{i=1}^{c}\sum_{j=1}^{p}\gamma_{ijk}u_{ijk} = 1$, as seguintes igualdades são verdadeiras:

$$\sum_{a=1}^{c}\sum_{b=1}^{p}\gamma_{abk}u_{abk} = \left(-\frac{\mu}{m}\right)^{-\frac{1}{m+1}} \sum_{a=1}^{c}\sum_{b=1}^{p}\left(\frac{1}{(u_{abk})^m d_{abk}}\right)^{-\frac{1}{m+1}} u_{abk} = 1 \tag{A.19}$$

A partir da segunda igualdade, é possível encontrar a seguinte equação:

$$\left(-\frac{\mu}{m}\right)^{-\frac{1}{m+1}} = \frac{1}{\sum_{a=1}^{c}\sum_{b=1}^{p}\left[(u_{abk})^m d_{abk}\right]^{\frac{1}{m+1}} u_{abk}} \tag{A.20}$$

Substituindo a expressão A.20 na expressão A.18, tem-se:

$$\gamma_{ijk} = \frac{1}{\sum_{a=1}^{c}\sum_{b=1}^{p}\left[(u_{abk})^m d_{abk}\right]^{\frac{1}{m+1}} u_{abk}} \left[(u_{ijk})^m d_{ijk}\right]^{\frac{1}{m+1}} \tag{A.21}$$

Finalmente:

$$\gamma_{ijk} = \frac{\left[(u_{ijk})^m d_{ijk}\right]^{\frac{1}{m+1}}}{\sum_{a=1}^{c}\sum_{b=1}^{p}\left[(u_{abk})^m d_{abk}\right]^{\frac{1}{m+1}} u_{abk}}.\square \tag{A.22}$$

A.6 Prova da Proposição 3.2.8

A série $\omega_t = J^2(\nu_t)$ decresce a cada iteração e converge.

Prova. De acordo com o algoritmo, as seguintes desigualdades (I), (II) e (III) acontece a cada iteração:

$$J^2(\mathbf{Y}^t, \mathbf{U}^t, \mathbf{\Gamma}^t) \overset{(I)}{\geq} J^2(\mathbf{Y}^{t+1}, \mathbf{U}^t, \mathbf{\Gamma}^t) \overset{(II)}{\geq} J^2(\mathbf{Y}^{t+1}, \mathbf{U}^{t+1}, \mathbf{\Gamma}^t) \overset{(III)}{\geq} J^2(\mathbf{Y}^{t+1}, \mathbf{U}^{t+1}, \mathbf{\Gamma}^{t+1}). \tag{A.23}$$

Seja $d(x_{kj}, y_{ij}^{(t)}) = (x_{kj} - y_{ij}^{(t)})^2$ a distância entre o objeto x_{kj} e o protótipo y_{ij} na iteração t. A desigualdade (I) acontece porque

$$J^2(\mathbf{Y}^t, \mathbf{U}^t, \mathbf{\Gamma}^t) = \sum_{i=1}^{c}\sum_{j=1}^{p}\gamma_{ij}^{(t)}\sum_{k=1}^{n}(u_{ijk}^{(t)})^m d(x_{kj}, y_{ij}^{(t)}),$$

$$J^2(\mathbf{Y}^{t+1}, \mathbf{U}^t, \mathbf{\Gamma}^t) = \sum_{i=1}^{c}\sum_{j=1}^{p}\gamma_{ij}^{(t)}\sum_{k=1}^{n}(u_{ijk}^{(t)})^m d(x_{kj}, y_{ij}^{(t+1)}),$$

e de acordo com a Proposição 3.2.5,

$$y_{ij}^{(t+1)} = \underbrace{argmin}_{y\in\Re}\gamma_{ij}^{(t)}\sum_{k=1}^{n}(u_{ijk}^{(t)})^m d(x_{kj}, y).$$

Além disso, a desigualdade (II) também acontece porque

$$J^2(\mathbf{Y}^{t+1}, \mathbf{U}^{t+1}, \mathbf{\Gamma}^t) = \sum_{i=1}^{c}\sum_{j=1}^{p}\gamma_{ij}^{(t)}\sum_{k=1}^{n}(u_{ijk}^{(t+1)})^m d(x_{kj}, y_{ij}^{(t+1)}),$$

e de acordo com a Proposição 3.2.6,

$$u_{ijk}^{(t+1)} = \underbrace{argmin}_{u\in[0,1]}\gamma_{ij}^{(t)}\sum_{k=1}^{n}(u)^m d(x_{kj}, y_{ij}^{(t+1)}).$$

E a desigualdade (III) também acontece porque

$$J^2(\mathbf{Y}^{t+1}, \mathbf{U}^{t+1}, \mathbf{\Lambda}^{t+1}) = \sum_{i=1}^{c}\sum_{j=1}^{p}\gamma_{ij}^{(t+1)}\sum_{k=1}^{n}(u_{ijk}^{(t+1)})^m d(x_{kj}, y_{ij}^{(t+1)}),$$

e de acordo com a Proposição 3.2.7,

$$\gamma_{ij}^{(t+1)} = \underbrace{argmin}_{\gamma\in\Re_+^*}\gamma\sum_{k=1}^{n}(u_{ijk}^{(t+1)})^m d(x_{kj}, y_{ij}^{(t+1)}).$$

Finalmente, uma vez que a série ω_t decresce e é limitado por $(J^3(v_t) \geq 0)$, ela converge. □

A.7 Prova da Proposição 3.2.9

A série $v_t = (\mathbf{Y}^t, \mathbf{U}^t, \mathbf{\Gamma}^t)$ converge.

Prova. Assuma que o ponto estacionário da série v_t é atingido na iteração $t = T$. Desta forma,

tem-se que $v_T = v_{T+1}$ e então $J^2(v_T) = J^2(v_{T+1})$, isto é, $J^2(\mathbf{Y}^T, \mathbf{U}^T, \mathbf{\Gamma}^T) = J^2(\mathbf{Y}^{T+1}, \mathbf{U}^{T+1}, \mathbf{\Gamma}^{T+1})$. A partir desta igualdade, a Proposição 3.2.8 pode ser reescrita como as igualdades (I), (II) e (III):

$$J^2(\mathbf{Y}^T, \mathbf{U}^T, \mathbf{\Gamma}^T) \overset{(I)}{=} J^2(\mathbf{Y}^{T+1}, \mathbf{U}^T, \mathbf{\Gamma}^T) \overset{(II)}{=} J^2(\mathbf{Y}^{T+1}, \mathbf{U}^{T+1}, \mathbf{\Gamma}^T) \overset{(III)}{=} J^2(\mathbf{Y}^{T+1}, \mathbf{U}^{T+1}, \mathbf{\Gamma}^{T+1}). \tag{A.24}$$

Da primeira igualdade (I), $\mathbf{Y}^T = \mathbf{Y}^{T+1}$ porque $\mathbf{Y}$ é único minimizando J^2 quando $\mathbf{U}^T$ e $\mathbf{\Gamma}^T$ são fixos. Da segunda igualdade (II), $\mathbf{U}^T = \mathbf{U}^{T+1}$ porque $\mathbf{U}$ é único minimizando J^2 quando $\mathbf{Y}^{T+1}$ e $\mathbf{\Gamma}^T$ são fixos. Além disso, a partir da terceira igualdade (III), $\mathbf{\Gamma}^T = \mathbf{\Gamma}^{T+1}$ porque $\mathbf{\Gamma}$ é único minimizando J^2 quando $\mathbf{Y}^{T+1}$ e $\mathbf{U}^{T+1}$ são fixos.

Finalmente, conclui-se que $v_T = v_{T+1}$. Esta conclusão acontece para todo $t \geq T$ e $v_t = v_T, \forall t \geq T$ e segue-se que v_t converge. □

A.8 Prova da Proposição 3.2.12

Fixando o protótipo y_{ij} e o grau de pertinência u_{ijk}, o peso λ_{ij} que minimiza o critério J^3 é atualizado usando a seguinte expressão sob a restrição $\prod_{h=1}^{p} \lambda_{ih} = 1$:

$$\lambda_{ij} = \frac{\left\{\prod_{h=1}^{p} \left[\sum_{k=1}^{n} (u_{ihk})^m d_{ihk}\right]\right\}^{\frac{1}{p}}}{\sum_{k=1}^{n} (u_{ijk})^m d_{ijk}}. \tag{A.25}$$

Prova. O peso pode ser calculado aplicando o método dos multiplicadores de Lagrange:

$$F_{ij}(\lambda) = \lambda_{ij} \sum_{k=1}^{n} (u_{ijk})^m d_{ijk} - \mu \left(\prod_{h=1}^{p} \lambda_{ih} - 1 \right). \tag{A.26}$$

F_{ij} é estacionário quando:

$$\frac{\partial F_{ij}}{\partial \lambda_{ij}} = \sum_{k=1}^{n} (u_{ijk})^m d_{ijk} - \mu \left(\frac{\prod_{h=1}^{p} \lambda_{ih}}{\lambda_{ij}} \right) = 0. \tag{A.27}$$

Da Equação A.27, obtem-se:

$$\sum_{k=1}^{n} (u_{ijk})^m d_{ijk} = \mu \left(\frac{\prod_{h=1}^{p} \lambda_{ih}}{\lambda_{ij}} \right) \tag{A.28}$$

Uma vez que a restrição é $\prod_{h=1}^{p} \lambda_{ih} = 1$, então:

$$\lambda_{ij} = \frac{\mu}{\sum_{k=1}^{n} (u_{ijk})^m d_{ijk}}. \tag{A.29}$$

De acordo com a restrição e usando a Equação A.29, as seguintes sentenças também são verdadeiras:

$$\prod_{h=1}^{p}\lambda_{ih}=\prod_{h=1}^{p}\left[\frac{\mu}{\sum_{k=1}^{n}(u_{ihk})^{m}d_{ihk}}\right]=\frac{\mu^{p}}{\prod_{h=1}^{p}\left[\sum_{k=1}^{n}(u_{ihk})^{m}d_{ihk}\right]}=1\Rightarrow$$

$$\mu=\left\{\prod_{h=1}^{p}\left[\sum_{k=1}^{n}(u_{ihk})^{m}d_{ihk}\right]\right\}^{\frac{1}{p}}. \quad \text{(A.30)}$$

Substituindo a Equação A.30 na Equação A.29, tem-se:

$$\lambda_{ij}=\frac{\left\{\prod_{h=1}^{p}\left[\sum_{k=1}^{n}(u_{ihk})^{m}d_{ihk}\right]\right\}^{\frac{1}{p}}}{\sum_{k=1}^{n}(u_{ijk})^{m}d_{ijk}}.\square \quad \text{(A.31)}$$

A.9 Prova da Proposição 3.2.17

Fixo o protótipo y_{ij} e o grau de pertinência u_{ijk}, o grau de possibilidade t_{ijk} que minimiza o critério J^4 é atualizado usando a seguinte expressão:

$$t_{ijk}=\left[1+\left(b\frac{d_{ijk}}{\pi_{ij}}\right)^{\frac{1}{(\eta-1)}}\right]^{-1}. \quad \text{(A.32)}$$

Prova. O critério J^4 sendo aditivo, o problema torna-se minimizar J^4_{ijk} diferenciando com respeito a t_{ijk}:

$$\frac{\partial J^4_{ijk}}{\partial t_{ijk}}=\eta b(t_{ijk})^{\eta-1}d_{ijk}+(-1)\eta\pi_{ij}(1-t_{ijk})^{\eta-1}=0, \quad \text{(A.33)}$$

O que é equivalente a:

$$(bd_{ijk})^{\frac{1}{(\eta-1)}}t_{ijk}-(\pi_{ij})^{\frac{1}{(\eta-1)}}+(\pi_{ij})^{\frac{1}{(\eta-1)}}t_{ijk}=0, \quad \text{(A.34)}$$

$$\Rightarrow t_{ijk}\left[(bd_{ijk})^{\frac{1}{(\eta-1)}}+(\pi_{ij})^{\frac{1}{(\eta-1)}}\right]=(\pi_{ij})^{\frac{1}{(\eta-1)}}, \quad \text{(A.35)}$$

$$\Rightarrow t_{ijk}=\frac{(\pi_{ij})^{\frac{1}{(\eta-1)}}}{(\pi_{ij})^{\frac{1}{(\eta-1)}}+(bd_{ijk})^{\frac{1}{(\eta-1)}}}, \quad \text{(A.36)}$$

Concluindo, tem-se que:

$$t_{ijk} = \left[1 + \left(b\frac{d_{ijk}}{\pi_{ij}}\right)^{\frac{1}{(\eta-1)}}\right]^{-1}.\square \tag{A.37}$$

A.10 Prova da Proposição 3.2.23

Fixando o protótipo y_{ij}, o grau de pertinência u_{ijk}e o grau de possibilidade t_{ijk}, o peso λ_{ij} que minimiza o critério J^5 é atualizado usando a seguinte expressão sob a restrição $\prod_{h=1}^{p}\lambda_{ih} = 1$:

$$\lambda_{ij} = \frac{\left\{\prod_{h=1}^{p}\left[\sum_{k=1}^{n}\left\{\left[a(u_{ihk})^m + b(t_{ihk})^\eta\right]d_{ihk} + \pi_{ih}(1-t_{ihk})^\eta\right\}\right]\right\}^{\frac{1}{p}}}{\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m + b(t_{ijk})^\eta\right]d_{ijk} + \pi_{ij}(1-t_{ijk})^\eta\right\}}. \tag{A.38}$$

Prova. O peso pode ser calculado aplicando o método dos multiplicadores de Lagrange:

$$F_{ij}(\lambda) = \lambda_{ij}\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m + b(t_{ijk})^\eta\right]d_{ijk} + \pi_{ij}(1-t_{ijk})^\eta\right\} - \mu\left(\prod_{h=1}^{p}\lambda_{ih} - 1\right). \tag{A.39}$$

F_{ij} é estacionário quando:

$$\frac{\partial F_{ij}}{\partial \lambda_{ij}} = \sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m + b(t_{ijk})^\eta\right]d_{ijk} + \pi_{ij}(1-t_{ijk})^\eta\right\} - \mu\left(\frac{\prod_{h=1}^{p}\lambda_{ih}}{\lambda_{ij}}\right) = 0. \tag{A.40}$$

Da Equação A.40, obtem-se:

$$\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m + b(t_{ijk})^\eta\right]d_{ijk} + \pi_{ij}(1-t_{ijk})^\eta\right\} = \mu\left(\frac{\prod_{h=1}^{p}\lambda_{ih}}{\lambda_{ij}}\right) \tag{A.41}$$

Uma vez que a restrição é $\prod_{h=1}^{p}\lambda_{ih} = 1$, então:

$$\lambda_{ij} = \frac{\mu}{\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^m + b(t_{ijk})^\eta\right]d_{ijk} + \pi_{ij}(1-t_{ijk})^\eta\right\}}. \tag{A.42}$$

De acordo com a restrição e usando a Equação A.42, as seguintes sentenças também são verdadeiras:

$$\prod_{h=1}^{p}\lambda_{ih}=\prod_{h=1}^{p}\frac{\mu}{\left\{\sum_{k=1}^{n}\left\{\left[a(u_{ihk})^{m}+b(t_{ihk})^{\eta}\right]d_{ihk}+\pi_{ih}(1-t_{ihk})^{\eta}\right\}\right\}}=$$

$$\frac{\mu^{p}}{\prod_{h=1}^{p}\left[\sum_{k=1}^{n}\left\{\left[a(u_{ihk})^{m}+b(t_{ihk})^{\eta}\right]d_{ihk}+\pi_{ih}(1-t_{ihk})^{\eta}\right\}\right]}=1\Rightarrow$$

$$\mu=\left\{\prod_{h=1}^{p}\left[\sum_{k=1}^{n}\left\{\left[a(u_{ihk})^{m}+b(t_{ihk})^{\eta}\right]d_{ihk}+\pi_{ih}(1-t_{ihk})^{\eta}\right\}\right]\right\}^{\frac{1}{p}}. \quad \text{(A.43)}$$

Substituindo a Equação A.43 na Equação A.42, tem-se:

$$\lambda_{ij}=\frac{\left\{\prod_{h=1}^{p}\left[\sum_{k=1}^{n}\left\{\left[a(u_{ihk})^{m}+b(t_{ihk})^{\eta}\right]d_{ihk}+\pi_{ih}(1-t_{ihk})^{\eta}\right\}\right]\right\}^{\frac{1}{p}}}{\sum_{k=1}^{n}\left\{\left[a(u_{ijk})^{m}+b(t_{ijk})^{\eta}\right]d_{ijk}+\pi_{ij}(1-t_{ijk})^{\eta}\right\}}.\square \quad \text{(A.44)}$$

A.11 Prova da Proposição 3.3.2

As seguintes relações acontecem para todo grupo $C_i(i=1,\ldots,c)$, variável $j(j=1,\ldots,p)$ e método multivariado$l$ $(l=1,2,3,4,5)$:

$$T^{l}=J^{l}+B^{l},\ T_{i}^{l}=J_{i}^{l}+B_{i}^{l},\ T_{j}^{l}=J_{j}^{l}+B_{j}^{l}\text{ and }T_{ij}^{l}=J_{ij}^{l}+B_{ij}^{l}. \quad \text{(A.45)}$$

Prova. Pode-se começar mostrando que $T^{1}=J^{1}+B^{1}$ acontece. Relembrando:

$$T^{1}=\sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk})^{m}(x_{kj}-y_{j})^{2}. \quad \text{(A.46)}$$

A distância pode ser reescrita como:

$$\begin{aligned}(x_{kj}-y_{j})^{2}&=\left[(x_{kj}-y_{ij})+(y_{ij}-y_{j})\right]^{2}\\&=(x_{kj}-y_{ij})^{2}+2(x_{kj}-y_{ij})(y_{ij}-y_{j})+(y_{ij}-y_{j})^{2}\end{aligned} \quad \text{(A.47)}$$

Desta forma, obtem-se:

$$T^{1}=\sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk})^{m}\left[(x_{kj}-y_{ij})^{2}+2(x_{kj}-y_{ij})(y_{ij}-y_{j})+(y_{ij}-y_{j})^{2}\right] \quad \text{(A.48)}$$

Isto é:

$$\begin{aligned}T^{1}=&\sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk})^{m}(x_{kj}-y_{ij})^{2}+2\sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk})^{m}(x_{kj}-y_{ij})(y_{ij}-y_{j})\\&+\sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk})^{m}(y_{ij}-y_{j})^{2}\end{aligned} \quad \text{(A.49)}$$

Uma vez que $J^1 = \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk})^m(x_{kj}-y_{ij})^2$ e $B^1 = \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk})^m(y_{ij}-y_j)^2$, assim pode-se substituir essas expressões na Equação A.49:

$$T^1 = J^1 + 2\sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk})^m(x_{kj}-y_{ij})(y_{ij}-y_j) + B^1 \tag{A.50}$$

Tem-se que $(x_{kj}-y_{ij})(y_{ij}-y_j) = y_{ij}(x_{kj}-y_{ij}) - y_j(x_{kj}-y_{ij})$. Assim:

$$\begin{aligned}\sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk})^m(x_{kj}-y_{ij})(y_{ij}-y_j) &= \sum_{i=1}^{c}\sum_{j=1}^{p}\sum_{k=1}^{n}(u_{ijk})^m\left[y_{ij}(x_{kj}-y_{ij}) - y_j(x_{kj}-y_{ij})\right] \\ &= \sum_{i=1}^{c}\sum_{j=1}^{p}\left\{y_{ij}\left[\sum_{k=1}^{n}(u_{ijk})^m x_{kj} - y_{ij}\sum_{k=1}^{n}(u_{ijk})^m\right]\right. \\ &\qquad \left. - y_j\left[\sum_{k=1}^{n}(u_{ijk})^m x_{kj} - y_{ij}\sum_{k=1}^{n}(u_{ijk})^m\right]\right\}\end{aligned} \tag{A.51}$$

A partir da definição do protótipo y_{ij}:

$$y_{ij} = \frac{\sum_{k=1}^{n}(u_{ijk})^m x_{kj}}{\sum_{k=1}^{n}(u_{ijk})^m} \Rightarrow \sum_{k=1}^{n}(u_{ijk})^m x_{kj} = y_{ij}\sum_{k=1}^{n}(u_{ijk})^m \Rightarrow$$

$$\sum_{k=1}^{n}(u_{ijk})^m x_{kj} - y_{ij}\sum_{k=1}^{n}(u_{ijk})^m = 0, \quad \text{A.52}$$

a seguinte expressão pode ser obtida:

$$\sum_{i=1}^{c}\sum_{j=1}^{p}\left\{y_{ij}\left[\sum_{k=1}^{n}(u_{ijk})^m x_{kj} - y_{ij}\sum_{k=1}^{n}(u_{ijk})^m\right] - y_j\left[\sum_{k=1}^{n}(u_{ijk})^m x_{kj} - y_{ij}\sum_{k=1}^{n}(u_{ijk})^m\right]\right\} = 0, \tag{A.53}$$

o que leva a concluir que $T^1 = J^1 + B^1$. As outras expressões podem ser facilmente obtidas de uma forma similar. □

APÊNDICE B - TESTES

Neste apêndice, serão apresentadas as tabelas contendo os *p-values* dos testes referentes aos experimentos realizados neste trabalho. Inicialmente, o Teste de Friedman (Friedman, 1937) é aplicado para todos os conjuntos de dados. É um teste não-paramétrico, isto é, não precisa que os dados assumam Distribuição Normal e, no contexto deste trabalho, verifica se é provável que os índices difusos obtidos pelos métodos venham da mesma população para um determinado conjunto de dados. Caso a hipótese nula do teste de Friedman seja rejeitada, serão realizados testes de hipótese para verificar diferenças entre pares de métodos. Para testar a alegação que a mediana dos valores do índice difuso encontrado por um método multivariado é igual à mediana correspondente a outro método de agrupamento, o teste de Wilcoxon-Mann-Whitney (Mann e Whitney, 1967) foi realizado. Este teste é não-paramétrico e é aplicado para duas amostras independentes. É quase tão eficaz quanto o teste *t-Student* em distribuições normais. O nível de significância adotado para esse teste é de 5%. Aqui, cada mediana associada a um método é escrita da seguinte forma: $\mu_l (l = 1, \ldots, 11)$, onde l é o índice de cada método — ver Tabela B.1.

GK	FCM	FCM-C1	PFCM	FCM-C2	W-FCM	MFCM	MFCM-G	MFCM-D	MPFCM	MPFCM-D
μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	μ_9	μ_{10}	μ_{11}

Tabela B.1: Métodos e medianas.

B.1 Testes de Friedman

Para verificar se os índices difusos calculados a partir dos resuldos dos métodos vêm da mesma população, testes de Friedman foram realizados. As tabelas B.2 a B.8 mostram os *p-values* dos testes para cada configuração.

Conf.	1	2	3	4	5	6	7	8	9	10
p-value	$8,15e-102$	$8,07e-97$	$2,61e-101$	$2,94e-102$	$3,24e-78$	$4,25e-89$	$7,25e-79$	$1,26e-101$	$1,84e-102$	$2,02e-101$

Tabela B.2: Teste de Friedman para configurações 1-10.

Conf.	0º	5º	10º	15º	20º	25º	30º	35º	40º	45º
p-value	$3,83e-103$	$3,16e-101$	$2,50e-102$	$7,63e-104$	$6,21e-103$	$6,80e-102$	$1,95e-96$	$5,01e-89$	$8,95e-86$	$1,81e-84$

Tabela B.3: Teste de Friedman para Configuração 11: variações no ângulo.

Conf.	0%	5%	10%	15%	20%	25%	30%
p-value	$3,62e-185$	$8,47e-128$	$1,24e-119$	$3,51e-119$	$8,93e-159$	$7,62e-187$	$3,04e-190$

Tabela B.4: Teste de Friedman para Configuração 12: variações no percentual de ruído.

Conf.	0%	5%	10%	15%	20%	25%	30%
p-value	$8,20e-116$	$2,77e-67$	$3,20e-52$	$6,87e-84$	$4,80e-56$	$2,33e-182$	$7,94e-183$

Tabela B.5: Teste de Friedman para Configuração 13: variações no percentual de ruído.

Conf.	Abalone	Banknote	Blood	Breast	B. Tissue	Climate	Ecoli	Glass	Haberman	ImageSeg	Ionosp.	Iris
p-value	$5,43e-34$	$5,06e-35$	$6,65e-25$	$3,52e-35$	$1,23e-15$		$1,15e-22$	$2,53e-14$	$1,02e-10$	$3,24e-16$	$5,91e-14$	$8,44e-26$

Tabela B.6: Teste de Friedman para conjuntos de dados da UCI — parte 1.

Conf.	Parkinsons	Pima	QSAR	Seeds	Sonar	StatlogV	Thyroid	Vertebral	WDBC	Wine	WineQlt	Yeast
p-value	$2,45e-21$	$1,10e-07$	$1,51e-13$	$1,18e-27$	$3,24e-24$	$2,42e-13$	$2,74e-30$	$1,14e-13$	$2,47e-29$	$3,24e-23$	$5,42e-12$	$2,52e-24$

Tabela B.7: Teste de Friedman para conjuntos de dados da UCI — parte 2.

Conf.	Chowdary	Dryskjot	Gordon	Nutt2	Shipp	Alizadeh1	Chen	Khan
p-value	$2,42e-23$	$4,42e-13$	$2,42e-12$	$3,53e-14$	$5,52e-21$	$6,24e-17$	$1,44e-19$	$8,24e-15$

Tabela B.8: Teste de Friedman para conjuntos de genes.

Os testes apresentados nas tabelas B.2 a B.8 mostraram que a mediana dos índices difusos obtidos pelos métodos vêm de populações diferentes para todas as configurações. Isto é, para uma dada configuração, existe pelo menos um método cuja mediana dos índices difusos diferem dos demais.

B.2 Testes de Wilcoxon-Mann-Whitney

Nesta parte, os métodos foram testados de modo a verificar se existe diferença estatística entre a mediana dos índices difusos obtidos por métodos multivariados. Para todos os tipos de conjuntos de dados (sintéticos, reais ou de genes), os testes foram realizados buscando mostrar que a suposição de um método multivariado é de fato comprovada, em pelo menos uma configuração. Portanto, o método MFCM foi comparado com todos os outros não multivariados presentes na literatura; MFCM-G foi comparado com o MFCM; MFCM-D foi testado contra o MFCM-G; MPFCM foi comparado com o PFCM; e o MPFCM-D analisado com o MPFCM. Tabela B.9 resume a suposição de cada método e com quais métodos ele é comparado.

Método	Suposição	Comparação
MFCM	Dispersões intra-grupo diferentes	GK, FCM, FCM-C[1], PFCM, FCM-C[2] e W-FCM
MFCM-G	Sobreposição de grupos	MFCM
MFCM-D	Dispersões inter-grupos diferentes	MFCM-G
MPFCM	Dispersões intra-grupo diferentes e dados ruidosos	PFCM
MPFCM-D	Dispersões inter-grupos diferentes e dados ruidosos	MPFCM

Tabela B.9: Métodos multivariados e suas suposições a respeito da configuração dos conjuntos de dados.

B.2.1 Dados Sintéticos

H_0: $\mu_1 = \mu_7$	H_0: $\mu_2 = \mu_7$	H_0: $\mu_3 = \mu_7$	H_0: $\mu_4 = \mu_7$	H_0: $\mu_5 = \mu_7$	H_0: $\mu_6 = \mu_7$
H_1: $\mu_1 \neq \mu_7$	H_1: $\mu_2 \neq \mu_7$	H_1: $\mu_3 \neq \mu_7$	H_1: $\mu_4 \neq \mu_7$	H_1: $\mu_5 \neq \mu_7$	H_1: $\mu_6 \neq \mu_7$
$1,34e-13$	$2,42e-17$	$4,24e-12$	$3,42e-13$	$3,24e-13$	$4,24e-16$
(Rejeita H_0)	(Rejeita H_0)	(Rejeita H_0)	(Rejeita H_0)	(Rejeita H_0)	(Rejeita H_0)

Tabela B.10: Testes de Wilcoxon-Mann-Whitney para MFCM: Configuração 3.

H_0: $\mu_7 = \mu_8$
H_1: $\mu_7 \neq \mu_8$
$3,24e-18$
(Rejeita H_0)

Tabela B.11: Teste de Wilcoxon-Mann-Whitney para MFCM-G: Configuração 1.

H_0: $\mu_8 = \mu_9$
H_1: $\mu_8 \neq \mu_9$
$2,42e-16$
(Rejeita H_0)

Tabela B.12: Teste de Wilcoxon-Mann-Whitney para MFCM-D: Configuração 2.

H_0: $\mu_4 = \mu_{10}$
H_1: $\mu_4 \neq \mu_{10}$
$2,42e-37$
(Rejeita H_0)

Tabela B.13: Teste de Wilcoxon-Mann-Whitney para MPFCM: Configuração 12 (30%).

H_0: $\mu_{10} = \mu_{11}$
H_1: $\mu_{10} \neq \mu_{11}$
$0,00026$
(Rejeita H_0)

Tabela B.14: Teste de Wilcoxon-Mann-Whitney para MPFCM-D: Configuração 13 (30%).

Os testes apresentados nas tabelas B.10 a B.14 confirmam a hipótese de que os métodos multivariados foram superiores a outros métodos em pelo menos uma configuração de dados sintéticos. Mais especificamente, MFCM foi superior aos não multivariados segundo a Configuração 3, que apresenta classes com dispersões diferentes intra-grupo, mas similar dispersão inter-grupo; MFCM-G foi superior que sua versão não multivariada MFCM segundo a Configuração 1, a qual apresenta um mais alto grau de dificuldade para encontrar a partição; MFCM-D foi superior ao MFCM-G de acordo com a Configuração 2, a qual apresenta classes

de diferentes formas e tamanhos; MPFCM mostrou obter uma melhor qualidade de agrupamento em comparação a sua versão não multivariada PFCM para a Configuração 1 2, com dados ruidosos; e MPFCM-D foi superior a sua versão não ponderada MPFCM na Configuração 13, a qual possui dados ruidosos e classes com diferentes veriabilidades intra-grupo. Essas conclusões são similarmente obtidas nos conjuntos de dados reais da UCI e dados de genes.

B.2.2 Dados Reais da UCI

Conjuntos	Testes					
Abalone	H_0: $\mu_1 = \mu_7$ H_1: $\mu_1 \neq \mu_7$	H_0: $\mu_2 = \mu_7$ H_1: $\mu_2 \neq \mu_7$	H_0: $\mu_3 = \mu_7$ H_1: $\mu_3 \neq \mu_7$	H_0: $\mu_4 = \mu_7$ H_1: $\mu_4 \neq \mu_7$	H_0: $\mu_5 = \mu_7$ H_1: $\mu_5 \neq \mu_7$	H_0: $\mu_6 = \mu_7$ H_1: $\mu_6 \neq \mu_7$
	$2,42e-22$ (Rejeita H_0)	$1,35e-23$ (Rejeita H_0)	$4,63e-22$ (Rejeita H_0)	$3,52e-24$ (Rejeita H_0)	$6,35e-20$ (Rejeita H_0)	$5,52e-22$ (Rejeita H_0)
Banknote	H_0: $\mu_1 = \mu_7$ H_1: $\mu_1 \neq \mu_7$	H_0: $\mu_2 = \mu_7$ H_1: $\mu_2 \neq \mu_7$	H_0: $\mu_3 = \mu_7$ H_1: $\mu_3 \neq \mu_7$	H_0: $\mu_4 = \mu_7$ H_1: $\mu_4 \neq \mu_7$	H_0: $\mu_5 = \mu_7$ H_1: $\mu_5 \neq \mu_7$	H_0: $\mu_6 = \mu_7$ H_1: $\mu_6 \neq \mu_7$
	$3,24e-16$ (Rejeita H_0)	$4,25e-13$ (Rejeita H_0)	$5,35e-16$ (Rejeita H_0)	$2,35e-13$ (Rejeita H_0)	$4,25e-15$ (Rejeita H_0)	$5,35e-14$ (Rejeita H_0)
Climate	H_0: $\mu_1 = \mu_7$ H_1: $\mu_1 \neq \mu_7$	H_0: $\mu_2 = \mu_7$ H_1: $\mu_2 \neq \mu_7$	H_0: $\mu_3 = \mu_7$ H_1: $\mu_3 \neq \mu_7$	H_0: $\mu_4 = \mu_7$ H_1: $\mu_4 \neq \mu_7$	H_0: $\mu_5 = \mu_7$ H_1: $\mu_5 \neq \mu_7$	H_0: $\mu_6 = \mu_7$ H_1: $\mu_6 \neq \mu_7$
	$6,35e-26$ (Rejeita H_0)	$2,42e-18$ (Rejeita H_0)	$5,35e-25$ (Rejeita H_0)	$3,63e-17$ (Rejeita H_0)	$3,53e-32$ (Rejeita H_0)	$5,53e-18$ (Rejeita H_0)
Glass	H_0: $\mu_1 = \mu_7$ H_1: $\mu_1 \neq \mu_7$	H_0: $\mu_2 = \mu_7$ H_1: $\mu_2 \neq \mu_7$	H_0: $\mu_3 = \mu_7$ H_1: $\mu_3 \neq \mu_7$	H_0: $\mu_4 = \mu_7$ H_1: $\mu_4 \neq \mu_7$	H_0: $\mu_5 = \mu_7$ H_1: $\mu_5 \neq \mu_7$	H_0: $\mu_6 = \mu_7$ H_1: $\mu_6 \neq \mu_7$
	$0,0035$ (Rejeita H_0)	$0,0014$ (Rejeita H_0)	$0,0053$ (Rejeita H_0)	$0,0042$ (Rejeita H_0)	$0,0052$ (Rejeita H_0)	$0,0024$ (Rejeita H_0)
WDBC	H_0: $\mu_1 = \mu_7$ H_1: $\mu_1 \neq \mu_7$	H_0: $\mu_2 = \mu_7$ H_1: $\mu_2 \neq \mu_7$	H_0: $\mu_3 = \mu_7$ H_1: $\mu_3 \neq \mu_7$	H_0: $\mu_4 = \mu_7$ H_1: $\mu_4 \neq \mu_7$	H_0: $\mu_5 = \mu_7$ H_1: $\mu_5 \neq \mu_7$	H_0: $\mu_6 = \mu_7$ H_1: $\mu_6 \neq \mu_7$
	$2,52e-24$ (Rejeita H_0)	$4,53e-26$ (Rejeita H_0)	$6,63e-23$ (Rejeita H_0)	$7,53e-26$ (Rejeita H_0)	$5,35e-20$ (Rejeita H_0)	$7,24e-27$ (Rejeita H_0)
WineQlt	H_0: $\mu_1 = \mu_7$ H_1: $\mu_1 \neq \mu_7$	H_0: $\mu_2 = \mu_7$ H_1: $\mu_2 \neq \mu_7$	H_0: $\mu_3 = \mu_7$ H_1: $\mu_3 \neq \mu_7$	H_0: $\mu_4 = \mu_7$ H_1: $\mu_4 \neq \mu_7$	H_0: $\mu_5 = \mu_7$ H_1: $\mu_5 \neq \mu_7$	H_0: $\mu_6 = \mu_7$ H_1: $\mu_6 \neq \mu_7$
	$3,53e-5$ (Rejeita H_0)	$5,35e-5$ (Rejeita H_0)	$3,72e-4$ (Rejeita H_0)	$5,72e-6$ (Rejeita H_0)	$2,62e-6$ (Rejeita H_0)	$2,42e-5$ (Rejeita H_0)
Yeast	H_0: $\mu_1 = \mu_7$ H_1: $\mu_1 \neq \mu_7$	H_0: $\mu_2 = \mu_7$ H_1: $\mu_2 \neq \mu_7$	H_0: $\mu_3 = \mu_7$ H_1: $\mu_3 \neq \mu_7$	H_0: $\mu_4 = \mu_7$ H_1: $\mu_4 \neq \mu_7$	H_0: $\mu_5 = \mu_7$ H_1: $\mu_5 \neq \mu_7$	H_0: $\mu_6 = \mu_7$ H_1: $\mu_6 \neq \mu_7$
	$4,24e-12$ (Rejeita H_0)	$2,53e-10$ (Rejeita H_0)	$7,53e-19$ (Rejeita H_0)	$7,35e-11$ (Rejeita H_0)	$8,35e-17$ (Rejeita H_0)	$7,35e-13$ (Rejeita H_0)

Tabela B.15: Testes de Wilcoxon-Mann-Whitney para MFCM: dados reais da UCI.

Conjuntos	Testes	Conjuntos	Testes
Blood	H_0: $\mu_7 = \mu_8$ H_1: $\mu_7 \neq \mu_8$ $2,42e-27$ (Rejeita H_0)	Pima	H_0: $\mu_7 = \mu_8$ H_1: $\mu_7 \neq \mu_8$ $4,24e-12$ (Rejeita H_0)
Ecoli	H_0: $\mu_7 = \mu_8$ H_1: $\mu_7 \neq \mu_8$ $3,53e-15$ (Rejeita H_0)	QSAR	H_0: $\mu_7 = \mu_8$ H_1: $\mu_7 \neq \mu_8$ $3,24e-17$ (Rejeita H_0)
Haberman	H_0: $\mu_7 = \mu_8$ H_1: $\mu_7 \neq \mu_8$ $2,42e-17$ (Rejeita H_0)	StatlogV	H_0: $\mu_7 = \mu_8$ H_1: $\mu_7 \neq \mu_8$ $5,34e-7$ (Rejeita H_0)
ImageSeg	H_0: $\mu_7 = \mu_8$ H_1: $\mu_7 \neq \mu_8$ $5,24e-13$ (Rejeita H_0)	Vertebral	H_0: $\mu_7 = \mu_8$ H_1: $\mu_7 \neq \mu_8$ $1,43e-25$ (Rejeita H_0)

Tabela B.16: Testes de Wilcoxon-Mann-Whitney para MFCM-G: dados reais da UCI.

Conjuntos	Testes	Conjuntos	Testes
Abalone	H_0: $\mu_8 = \mu_9$ H_1: $\mu_8 \neq \mu_9$ $3,32e-21$ (Rejeita H_0)	Parkinsons	H_0: $\mu_8 = \mu_9$ H_1: $\mu_8 \neq \mu_9$ $2,42e-19$ (Rejeita H_0)
Breast	H_0: $\mu_8 = \mu_9$ H_1: $\mu_8 \neq \mu_9$ $2,52e-28$ (Rejeita H_0)	Seeds	H_0: $\mu_8 = \mu_9$ H_1: $\mu_8 \neq \mu_9$ $4,24e-25$ (Rejeita H_0)
BreastTissue	H_0: $\mu_8 = \mu_9$ H_1: $\mu_8 \neq \mu_9$ $6,34e-35$ (Rejeita H_0)	Sonar	H_0: $\mu_8 = \mu_9$ H_1: $\mu_8 \neq \mu_9$ $3,53e-42$ (Rejeita H_0)
Iris	H_0: $\mu_8 = \mu_9$ H_1: $\mu_8 \neq \mu_9$ $2,52e-19$ (Rejeita H_0)	Wine	H_0: $\mu_8 = \mu_9$ H_1: $\mu_8 \neq \mu_9$ $6,35e-17$ (Rejeita H_0)

Tabela B.17: Testes de Wilcoxon-Mann-Whitney para MFCM-D: dados reais da UCI.

Conjuntos	Testes	Conjuntos	Testes
Abalone	H_0: $\mu_4 = \mu_{10}$ H_1: $\mu_4 \neq \mu_{10}$ $2,52e-15$ (Rejeita H_0)	Parkinsons	H_0: $\mu_4 = \mu_{10}$ H_1: $\mu_4 \neq \mu_{10}$ $3,24e-17$ (Rejeita H_0)
B, Tissue	H_0: $\mu_4 = \mu_{10}$ H_1: $\mu_4 \neq \mu_{10}$ $3,53e-24$ (Rejeita H_0)	Sonar	H_0: $\mu_4 = \mu_{10}$ H_1: $\mu_4 \neq \mu_{10}$ $2,53e-9$ (Rejeita H_0)
Climate	H_0: $\mu_4 = \mu_{10}$ H_1: $\mu_4 \neq \mu_{10}$ $4,52e-12$ (Rejeita H_0)	StatlogV	H_0: $\mu_4 = \mu_{10}$ H_1: $\mu_4 \neq \mu_{10}$ $3,53e-7$ (Rejeita H_0)
Iris	H_0: $\mu_4 = \mu_{10}$ H_1: $\mu_4 \neq \mu_{10}$ $3,63e-9$ (Rejeita H_0)	Wine	H_0: $\mu_4 = \mu_{10}$ H_1: $\mu_4 \neq \mu_{10}$ $6,52e-13$ (Rejeita H_0)

Tabela B.18: Testes de Wilcoxon-Mann-Whitney para MPFCM: dados reais da UCI.

Conjuntos	Testes	Conjuntos	Testes
Abalone	H_0: $\mu_{10} = \mu_{11}$ H_1: $\mu_{10} \neq \mu_{11}$ $0,0035$ (Rejeita H_0)	StatlogV	H_0: $\mu_{10} = \mu_{11}$ H_1: $\mu_{10} \neq \mu_{11}$ $3,53e-7$ (Rejeita H_0)
Banknote	H_0: $\mu_{10} = \mu_{11}$ H_1: $\mu_{10} \neq \mu_{11}$ $0,0001$ (Rejeita H_0)	Thyroid	H_0: $\mu_{10} = \mu_{11}$ H_1: $\mu_{10} \neq \mu_{11}$ $2,53e-9$ (Rejeita H_0)
Breast	H_0: $\mu_{10} = \mu_{11}$ H_1: $\mu_{10} \neq \mu_{11}$ $0,0042$ (Rejeita H_0)	WDBC	H_0: $\mu_{10} = \mu_{11}$ H_1: $\mu_{10} \neq \mu_{11}$ $0,0002$ (Rejeita H_0)
Climate	H_0: $\mu_{10} = \mu_{11}$ H_1: $\mu_{10} \neq \mu_{11}$ $2,42e-5$ (Rejeita H_0)	WineQlt	H_0: $\mu_{10} = \mu_{11}$ H_1: $\mu_{10} \neq \mu_{11}$ $2,52e-5$ (Rejeita H_0)

Tabela B.19: Testes de Wilcoxon-Mann-Whitney para MPFCM-D: dados reais da UCI.

Os testes apresentados nas tabelas B.15 a B.19 suportam a hipótese de que os métodos multivariados obtiveram uma melhor qualidedade de agrupamento em comparação aos outros métodos em diversas configurações de dados reais da UCI.

B.2.3 Dados de Genes

Conjuntos	Testes					
Chowdary	H_0: $\mu_1 = \mu_7$ H_1: $\mu_1 \neq \mu_7$	H_0: $\mu_2 = \mu_7$ H_1: $\mu_2 \neq \mu_7$	H_0: $\mu_3 = \mu_7$ H_1: $\mu_3 \neq \mu_7$	H_0: $\mu_4 = \mu_7$ H_1: $\mu_4 \neq \mu_7$	H_0: $\mu_5 = \mu_7$ H_1: $\mu_5 \neq \mu_7$	H_0: $\mu_6 = \mu_7$ H_1: $\mu_6 \neq \mu_7$
	$3,24e-24$ (Rejeita H_0)	$4,24e-32$ (Rejeita H_0)	$5,24e-26$ (Rejeita H_0)	$5,13e-31$ (Rejeita H_0)	$2,42e-46$ (Rejeita H_0)	$6,53e-39$ (Rejeita H_0)
Dryskjot	H_0: $\mu_1 = \mu_7$ H_1: $\mu_1 \neq \mu_7$	H_0: $\mu_2 = \mu_7$ H_1: $\mu_2 \neq \mu_7$	H_0: $\mu_3 = \mu_7$ H_1: $\mu_3 \neq \mu_7$	H_0: $\mu_4 = \mu_7$ H_1: $\mu_4 \neq \mu_7$	H_0: $\mu_5 = \mu_7$ H_1: $\mu_5 \neq \mu_7$	H_0: $\mu_6 = \mu_7$ H_1: $\mu_6 \neq \mu_7$
	$3,24e-18$ (Rejeita H_0)	$4,22e-23$ (Rejeita H_0)	$3,42e-19$ (Rejeita H_0)	$5,52e-32$ (Rejeita H_0)	$3,24e-34$ (Rejeita H_0)	$5,42e-17$ (Rejeita H_0)
Alizadeh1	H_0: $\mu_1 = \mu_7$ H_1: $\mu_1 \neq \mu_7$	H_0: $\mu_2 = \mu_7$ H_1: $\mu_2 \neq \mu_7$	H_0: $\mu_3 = \mu_7$ H_1: $\mu_3 \neq \mu_7$	H_0: $\mu_4 = \mu_7$ H_1: $\mu_4 \neq \mu_7$	H_0: $\mu_5 = \mu_7$ H_1: $\mu_5 \neq \mu_7$	H_0: $\mu_6 = \mu_7$ H_1: $\mu_6 \neq \mu_7$
	$5,63e-33$ (Rejeita H_0)	$4,24e-35$ (Rejeita H_0)	$5,22e-34$ (Rejeita H_0)	$4,52e-35$ (Rejeita H_0)	$1,74e-34$ (Rejeita H_0)	$1,42e-36$ (Rejeita H_0)
Chen	H_0: $\mu_1 = \mu_7$ H_1: $\mu_1 \neq \mu_7$	H_0: $\mu_2 = \mu_7$ H_1: $\mu_2 \neq \mu_7$	H_0: $\mu_3 = \mu_7$ H_1: $\mu_3 \neq \mu_7$	H_0: $\mu_4 = \mu_7$ H_1: $\mu_4 \neq \mu_7$	H_0: $\mu_5 = \mu_7$ H_1: $\mu_5 \neq \mu_7$	H_0: $\mu_6 = \mu_7$ H_1: $\mu_6 \neq \mu_7$
	$5,24e-52$ (Rejeita H_0)	$6,52e-52$ (Rejeita H_0)	$2,52e-51$ (Rejeita H_0)	$2,52e-51$ (Rejeita H_0)	$2,52e-50$ (Rejeita H_0)	$2,42e-53$ (Rejeita H_0)

Tabela B.20: Testes de Wilcoxon-Mann-Whitney para MFCM: dados de gene.

Conjuntos	Testes	Conjuntos	Testes
Gordon	H_0: $\mu_7 = \mu_8$ H_1: $\mu_7 \neq \mu_8$	Khan	H_0: $\mu_7 = \mu_8$ H_1: $\mu_7 \neq \mu_8$
	$3,53e-48$ (Rejeita H_0)		$2,53e-46$ (Rejeita H_0)

Tabela B.21: Testes de Wilcoxon-Mann-Whitney para MFCM-G: dados de gene.

Conjuntos	Testes	Conjuntos	Testes
Dryskjot	H_0: $\mu_8 = \mu_9$ H_1: $\mu_8 \neq \mu_9$ $3,46e-47$ (Rejeita H_0)	Shipp	H_0: $\mu_8 = \mu_9$ H_1: $\mu_8 \neq \mu_9$ $1,43e-36$ (Rejeita H_0)

Tabela B.22: Testes de Wilcoxon-Mann-Whitney para MFCM-D: dados de gene.

Conjuntos	Testes	Conjuntos	Testes
Chowdary	H_0: $\mu_4 = \mu_{10}$ H_1: $\mu_4 \neq \mu_{10}$ $5,24e-37$ (Rejeita H_0)	Alizadeh1	H_0: $\mu_4 = \mu_{10}$ H_1: $\mu_4 \neq \mu_{10}$ $5,24e-58$ (Rejeita H_0)

Tabela B.23: Testes de Wilcoxon-Mann-Whitney para MPFCM: dados de gene.

Conjuntos	Testes	Conjuntos	Testes
Dryskjot	H_0: $\mu_{10} = \mu_{11}$ H_1: $\mu_{10} \neq \mu_{11}$ $3,45e-40$ (Rejeita H_0)	Alizadeh1	H_0: $\mu_{10} = \mu_{11}$ H_1: $\mu_{10} \neq \mu_{11}$ $4,14e-56$ (Rejeita H_0)

Tabela B.24: Testes de Wilcoxon-Mann-Whitney para MPFCM-D: dados de gene.

Tabelas B.20 a B.24 mostraram que, em pelo menos dois conjuntos de dados de genes, um método multivariado foi superior a outro.