



Universidade Federal da Paraíba
Centro de Informática
Programa de Pós-Graduação em Modelagem Matemática e Computacional

AGRUPAMENTO *FUZZY* NO ESPAÇO DE CARACTERÍSTICAS
BASEADO NO *KERNEL* DE MAHALANOBIS COM DISTÂNCIAS
QUADRÁTICAS ADAPTATIVAS

Evilasio Macedo Félix

João Pessoa
Dezembro de 2018



Universidade Federal da Paraíba
Centro de Informática
Programa de Pós-Graduação em Modelagem Matemática e Computacional

**AGRUPAMENTO *FUZZY* NO ESPAÇO DE CARACTERÍSTICAS
BASEADO NO *KERNEL* DE MAHALANOBIS COM DISTÂNCIAS
QUADRÁTICAS ADAPTATIVAS**

Evilasio Macedo Félix

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional, UFPB, da Universidade Federal da Paraíba, como parte dos requisitos necessários à obtenção do título de Mestre em Modelagem Matemática e Computacional.

Orientador: Marcelo Rodrigo Portela Ferreira

João Pessoa
Dezembro de 2018

Catálogo na publicação
Seção de Catalogação e Classificação

F316a Félix, Evilasio Macedo.

AGRUPAMENTO FUZZY NO ESPAÇO DE CARACTERÍSTICAS BASEADO
NO KERNEL DE MAHALANOBIS COM DISTÂNCIAS QUADRÁTICAS
ADAPTATIVAS / Evilasio Macedo Félix. - João Pessoa,
2018.

62 f. : il.

Orientação: Marcelo Rodrigo Portela Ferreira.
Dissertação (Mestrado) - UFPB/CI/PPGMMC.

1. Agrupamento fuzzy. 2. Kernel de Mahalanobis. 3.
Espaço de características. I. Ferreira, Marcelo Rodrigo
Portela. II. Título.

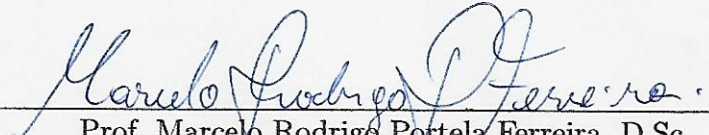
UFPB/BC

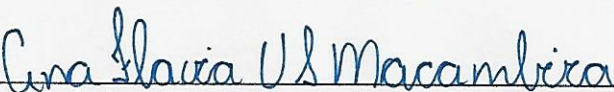
**AGRUPAMENTO *FUZZY* NO ESPAÇO DE CARACTERÍSTICAS
BASEADO NO *KERNEL* DE MAHALANOBIS COM DISTÂNCIAS
QUADRÁTICAS ADAPTATIVAS**

Evilasio Macedo Félix

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO PROGRAMA DE
PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL
(PPGMMC) DO CENTRO DE INFORMÁTICA DA UNIVERSIDADE FEDERAL
DA PARAÍBA COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A
OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM MODELAGEM
MATEMÁTICA E COMPUTACIONAL.

Examinada por:


Prof. Marcelo Rodrigo Portela Ferreira, D.Sc.


Prof. Ana Flavia Uzeda Dos Santos Macambira, D.Sc.


Prof. Sérgio de Carvalho Bezerra, D.Sc.

JOÃO PESSOA, PB – BRASIL
DEZEMBRO DE 2018

*A meus pais, Adão Félix e Maria
Macedo, meus irmãos, Ednael e
Charles, pessoas indispensáveis.*

Agradecimentos

A Deus

- Pela graça alcançada. Por iluminar meu caminho até aqui, dando força nos momentos difíceis e sabedoria para fazer as escolhas corretas.

A Família

- A meus pais, Adão Félix e Maria Macedo por todas as vezes que se sacrificaram para que eu e meus irmãos continuássemos a estudar, pelas orientações e broncas quando necessárias. Aos irmãos Ednael e Charles, pela ajuda em todos os momentos, pelos seus ensinamentos, por sempre estarem presentes em minhas perdas e conquistas. Sua alegria e carinho nos momentos mais difíceis animava meu coração e servia de estímulo para não desistir. A meus avós, Raimundo e Iraci pelo carinho, confiança e apoio que sempre demonstraram.

Aos Amigos

- Da República "Mangabeira Cite": Jairo, Vanlex, Aluísio, Iago e república "Vila do Chaves": Marlos e Wilson, pelos momentos de conversa, estudo e brincadeiras. Aos colegas de sala: Aline, Rita e Marlos que fizeram os momentos de estudo serem mais prazerosos. A Diego Alexandre, Erica Castro, Letícia Freitas, Ananias Félix e Padre Ademar por se fazer presente quando solicitado. A vocês deixo meus sinceros agradecimentos!

Aos Professores

- Ao meu orientador, Marcelo Rodrigo pelas orientações, atenção e paciência ao longo da realização deste trabalho. Como também, aos demais professores do PPGMMC, em especial, a professora Ana Paula Wyse pelo auxílio no decorrer do curso. Destaco ainda os professores, Sérgio de Carvalho Bezerra, Pedro Rafael Diniz Marinho e Ana Flavia Uzeda Dos Santos Macambira por terem aceitado, prontamente, participar da banca examinadora na certeza de colaborar com o estudo. A CAPES, pelo apoio financeiro. A todos que, de forma direta ou indireta, me ajudaram e contribuíram com o estudo, meus sinceros agradecimentos.

O freguês nunca se arrepende de fazer o que é certo !
Raimundo Macedo

Resumo da Dissertação apresentada ao PPGMMC/CI/UFPB como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

AGRUPAMENTO *FUZZY* NO ESPAÇO DE CARACTERÍSTICAS BASEADO NO *KERNEL* DE MAHALANOBIS COM DISTÂNCIAS QUADRÁTICAS ADAPTATIVAS

Evilasio Macedo Félix

Dezembro/2018

Orientador: Marcelo Rodrigo Portela Ferreira

Programa: Modelagem Matemática e Computacional

Apresenta-se, nesta dissertação de mestrado, métodos de agrupamento *fuzzy* no espaço de características baseado no *kernel* de Mahalanobis com distâncias quadráticas adaptativas, rotulados respectivamente por (KFCM.FS.GD, KFCM.FS.GF, KFCM.FS.LD e KFCM.FS.LF). Este estudo é uma extensão do trabalho desenvolvido em [26]. Os métodos propostos foram baseados no *kernel* de Mahalanobis a partir de distâncias quadráticas adaptativas definidas por matrizes de covariâncias simétricas positivas definidas. Estas matrizes de covariâncias são, diagonais e completas (não diagonais), comuns a todos os grupos e diferentes para cada grupo, determinadas sob o enfoque de agrupamento no espaço de características, que realiza um mapeamento de cada observação por meio de uma função não-linear Φ e então obtêm os centróides dos grupos no espaço de recursos. Esta técnica permite que ao passarmos para um espaço de mais alta dimensão (espaço de características), um conjunto de observações no espaço de entrada não-linearmente separável torna-se separável linearmente no espaço de características. Os algoritmos propostos foram comparados com os diversos métodos de agrupamento tradicionais conhecidos na literatura, como o *fuzzy k*-médias e suas versões baseadas no *kernel* Gaussiano, como também os métodos desenvolvido por [26]. A avaliação foi realizada através de experimentos numéricos com dados simulados e reais.

Palavras-chave: Agrupamento *fuzzy*. *Kernel* de Mahalanobis. Espaço de características.

Abstract of Dissertation presented to PPGMMC/CI/UFPB as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

***FUZZY CLUSTERING IN FEATURE SPACE BASED ON THE
MAHALANOBIS *KERNEL* WITH ADAPTIVE QUADRATIC
DISTANCES***

Evilasio Macedo Félix

December/2018

Advisor: Marcelo Rodrigo Portela Ferreira

Program: Computational Mathematical Modelling

In this master's thesis, fuzzy grouping methods are presented in the space of features based on the Mahalanobis kernel with adaptive quadratic distances, labeled respectively by (KFCM.FS.GD, KFCM.FS.GF, KFCM.FS.LD e KFCM.FS.LF). This study is an extension of the work developed in [26]. The proposed methods were based on the Mahalanobis *kernel* from adaptive quadratic distances defined by defined positive symmetric covariance matrices. These matrices of covariances are diagonal and complete (not diagonal), common to all groups and different for each group, determined under the clustering approach in the feature space, which performs a mapping of each observation by means of a nonlinear Φ and then obtain the centroids of the groups in the resource space. This technique allows that when we move to a space of higher dimension (space of characteristics), a set of observations in the non-linearly separable input space becomes linearly separable in the space of characteristics. The proposed algorithms were compared with the various traditional clustering methods known in the literature, such as *fuzzy k-means* and their versions based on the Gaussian *kernel*, as well as the methods developed by [26]. The evaluation was performed through numerical experiments with simulated and real data.

keyword: *Fuzzy* clustering. Mahalanobis *Kernel*. Feature space.

Sumário

Lista de Figuras	xi
Lista de Tabelas	xii
Lista de Símbolos	xiii
1 Introdução	1
1.1 Objetivos	5
1.2 Estrutura do Trabalho	6
2 Revisão Bibliográfica	7
2.1 Introdução	7
2.2 Métodos Particionais	7
2.2.1 Método <i>fuzzy c</i> -médias	8
2.2.2 Agrupamento baseado em <i>kernel</i>	10
2.2.3 <i>Kernel c</i> -médias difuso baseado em kernelização da métrica . .	12
2.2.4 <i>Kernel c</i> -médias difuso no espaço de características	13
3 Métodos Propostos	15
3.1 Considerações Iniciais quanto a Distância de Mahalanobis	15
3.2 Métodos de agrupamento <i>fuzzy</i> no espaço de características baseado no <i>kernel</i> de Mahalanobis com distâncias quadráticas adaptativas . .	17
4 Experimentos computacionais	28
4.1 Índices de Avaliação	28
4.2 Conjunto de Dados Simulados	30
4.3 Avaliação com Dados Simulados	32
4.4 Conjunto de Dados Reais	36
4.5 Avaliação com Dados Reais	39
5 Conclusões	43
5.1 Sugestões para Trabalhos Futuros	44

Lista de Figuras

1.1	(a) Satisfatório e (b) Não Satisfatório	3
1.2	Mapeamento baseado em função <i>kernel</i>	3
4.1	Conjunto de Dados Simulados	32

Lista de Tabelas

4.1	Matriz de confusão	29
4.2	Configurações dos cenários	31
4.3	Desempenho dos algoritmos de agrupamento <i>fuzzy</i> no conjunto de dados simulados 1: média e desvio-padrão (entre parênteses) dos índices CR e OERC.	33
4.4	Desempenho dos algoritmos de agrupamento <i>fuzzy</i> no conjunto de dados simulados 2: média e desvio-padrão (entre parênteses) dos índices CR e OERC.	34
4.5	Desempenho dos algoritmos de agrupamento <i>fuzzy</i> no conjunto de dados simulados 3: média e desvio-padrão (entre parênteses) dos índices CR e OERC.	35
4.6	Desempenho dos algoritmos de agrupamento <i>fuzzy</i> no conjunto de dados simulados 4: média e desvio-padrão (entre parênteses) dos índices CR e OERC.	35
4.7	Algoritmos de agrupamento <i>fuzzy</i> com melhor desempenho no conjunto de dados simulados 1,2,3 e 4.	36
4.8	Dimensões dos conjuntos de dados reais	37
4.9	Desempenho dos métodos de agrupamento <i>fuzzy</i> nos conjuntos de dados reais dos índices CR e OERC.	41
4.10	Algoritmos de agrupamento <i>fuzzy</i> com melhor desempenho no conjunto de dados simulados e reais.	42

Lista de Símbolos

X	Espaço original dos dados a serem agrupados, p. 3
$\mathcal{F}$	Espaço de características, p. 3
$\mathbf{x}_i$	Observação pertencente ao $\mathbb{R}^p$, p. 4
Φ	Função não-linear que mapeia os elementos de X em $\mathcal{F}$, p. 5
φ^2	Quadrado da distância euclidiana, p. 9
L	Função lagrangeana, p. 9
α	Multiplicador de lagrange, p. 9
U	Matriz difusa, p. 9
$\mathcal{K}$	<i>Kernel</i> de Mercer, p. 10
$\Phi(\mathbf{x}_i)$	Observação $\mathbf{x}_i$ mapeada para o espaço de características, p. 10
K	Matriz <i>kernel</i> , p. 11
$\mathbf{v}_k^\Phi$	Centróide do k -ésimo grupo no espaço de características, p. 11
Σ	Matriz de covariâncias, p. 16
$\mathcal{K}^{(M)}$	<i>Kernel</i> adaptativo de Mahalanobis, p. 17
ϕ	Medida de distância adequada, p. 18
d_M^2	Distância quadrática definida por uma matriz simétrica definida positiva $\mathbf{M}$, que é a mesma para todos os grupos, p. 19
$d_{M_k}^2$	Distância quadrática definida por uma matriz simétrica definida positiva $\mathbf{M}_k$, diferente para cada grupo, p. 20
H	Matriz hessiana, p. 28
$\mathcal{P}$	Partição <i>a priori</i> em c classes, p. 29

P	Partição em K grupos fornecidos por um algoritmo de agrupamento, p. 29
n_{ik}	Número de observações na classe $\mathcal{P}_i$ e no grupo P_k , p. 29
ℓ	Função de verossimilhança, p. 30
μ	Vetor de médias, p. 31

Capítulo 1

Introdução

O avanço tecnológico nas diversas áreas das ciências vem tornando o mundo cada vez mais globalizado, impondo a necessidade de informações cada vez mais precisas e direcionadas. Com esse avanço, vários setores da sociedade começaram a gerar e trabalhar com uma grande massa de dados no meio digital, nesta perspectiva os dados gerados contém informações importantes que não são fáceis de serem extraídas. Como consequência, surgiu a área de ciências de dados. Frente a este cenário, surge a necessidade de se analisar os dados automaticamente na perspectiva de extrair ou organizar melhor o conjunto de informações [21, 27]. Dentre as formas disponíveis de se analisarem os dados automaticamente, evidenciamos a forma análise de agrupamento que, procura encontrar grupos de dados semelhantes segundo alguma métrica.

Nesta perspectiva, o processo de agrupar consiste em colocar em um mesmo grupo, observações que sejam similares de acordo com algum critério pré-determinado. Este critério, baseia-se normalmente em uma função de dissimilaridade (ou similaridade). O agrupamento é tal que o grau de associação entre observações do mesmo grupo é alto e entre observações de grupos diferentes é baixo, [9]. Em outras palavras, todo agrupamento é feito com objetivo de maximizar a homogeneidade dentro de cada grupo e maximizar a heterogeneidade entre os grupos, i.e., elementos que fazem parte de um mesmo grupo devem apresentar alta similaridade (similaridade intra-grupo), mas devem ser muito dissimilares à elementos de outros grupos (similaridade inter-grupos).

Historicamente, o desenvolvimento da metodologia de agrupar tem sido um esforço verdadeiramente interdisciplinar. Taxonomistas, cientistas sociais, psicólogos, biólogos, estatísticos, matemáticos, engenheiros, cientistas da computação, pesquisadores médicos e outros profissionais que coletam e processam dados reais contribuíram na construção da metodologia de agrupamento [21].

Técnicas de agrupamento vêm sendo constantemente aplicadas em uma variedade de áreas da ciência, tais como, taxonomia, processamento de imagens, mineração de

dados, recuperação de informação, dentre outras. Exemplos de aplicações incluem: segmentação de mercado, segmentação de imagens, agrupar resultados de buscas na internet, agrupar páginas da internet com base em seu conteúdo, agrupar pessoas em redes sociais com base em suas preferências ou características. Em tais situações, os métodos de agrupamento comumente utilizados na literatura são divididos em duas grandes categorias, métodos hierárquicos e métodos particionais, cada um estruturado com diferentes tipos de algoritmos [23].

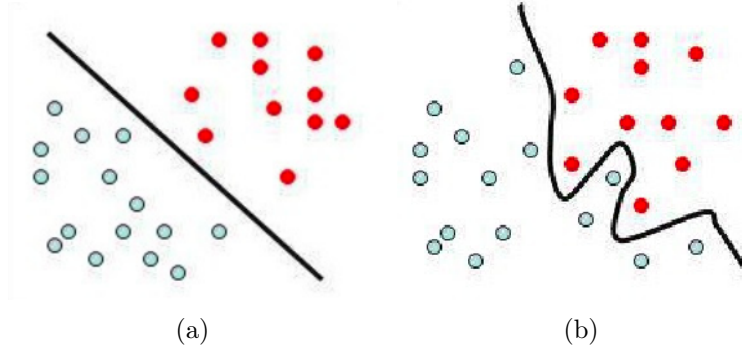
Os métodos hierárquicos, são técnicas que particionam os dados sucessivamente, determinando uma cadeia hierárquica de agrupamentos. Essa estrutura facilita a visualização do agrupamento em cada etapa do algoritmo, em especial na construção da sequência aninhada de partições do conjunto de observações de entrada. Nesta perspectiva, os métodos hierárquicos são formulados em duas versões: os aglomerativos e os divisivos. Os aglomerativos atuam inicialmente com cada observação formando seu próprio agrupamento e gradualmente os grupos são unidos [21, 23]. São exemplos de métodos hierárquicos aglomerativos: *dendrograma*, *Coefficiente Aglomerativo (CA)* e *Banner de Dissimilaridade*. Em contrapartida, os métodos hierárquicos divisivos começam com um único agrupamento formado por todas as observações e gradualmente dividem o agrupamento em agrupamentos menores, finalizando com um agrupamento por observação. Como exemplo de métodos hierárquicos divisivos, temos: *dendrograma*, *Coefficiente Divisivo (CD)* e *Banner de Dissimilaridade*, a descrição e funcionamento desses métodos aglomerativos e divisivos podem ser encontradas em [27].

Para os métodos de agrupamento particionais, a ideia é obter uma partição única do conjunto de observações de entrada em um número fixo de grupos, tipicamente através da otimização (geralmente local), de uma função objetivo. Estes métodos são estruturados sob dois pilares: agrupamento rígido (*hard*) e agrupamento difuso (*fuzzy*). Nos métodos de agrupamento rígido, cada observação do banco de dados é associada a um único grupo, isto significa que os grupos são naturalmente disjuntos e não se sobrepõem. Por outro lado, nos métodos de agrupamento do tipo difuso, foco desta pesquisa, uma observação pode pertencer a todos os grupos com um certo grau de pertinência. Uma exposição detalhada dos principais métodos de agrupamento difuso pode ser encontrada em [19].

Nos métodos de agrupamento particionais (rígido e difuso), as medidas de distância são exemplos básicos de medidas de dissimilaridade e a distância Euclidiana é a mais comumente utilizada. Segundo [12], métodos de agrupamento baseados na distância Euclidiana apresentam bom desempenho quando aplicados a conjuntos de dados nos quais os grupos são aproximadamente hiperesféricos e aproximadamente linearmente separáveis, ver Figura 1.1(a). Contudo, quando a estrutura dos dados é complexa, i.e., grupos com formas não hiperesféricas e/ou observações

não-linearmente separáveis, ver Figura 1.1(b), esses métodos podem não apresentar desempenho satisfatório.

Figura 1.1: (a) Satisfatório e (b) Não Satisfatório

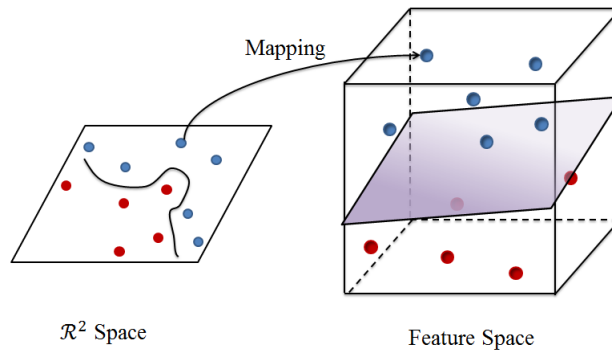


Fonte: Página O Estudante Independente ¹.

Por causa dessa limitação, diversos métodos capazes de lidar com dados cuja estrutura é complexa têm sido propostos, dentre os quais, métodos de agrupamento baseados em funções *kernel*.

A ideia dos métodos baseados em *kernel* envolve a realização de um mapeamento não-linear arbitrário Φ do espaço original (input space) p -dimensional $X \subset \mathbb{R}^p$ para um espaço de dimensão mais alta, chamado espaço de características (feature space), $\mathcal{F}$.

Figura 1.2: Mapeamento baseado em função *kernel*.



Fonte: Junior, 2010.

A razão para passarmos a dimensões mais altas é que em tais dimensões pode ser possível obter grupos bem definidos e linearmente separáveis. Métodos baseados em *kernel* possuem a vantagem de que produtos internos no espaço de características podem ser expressos por um *kernel* de Mercer $\mathcal{K}$, e cuja definição matemática encontra-se no Capítulo 2, Seção 2.2.2.

¹Disponível em: <<https://juliocprocha.wordpress.com/2017/07/27/perceptron-para-classificacao-passo-a-passo/>>. Acesso em: 27 out. 2018.

De acordo com a literatura, [14] desenvolveu um algoritmo que mapeia o espaço de entrada em um espaço de características e executa o agrupamento nesse novo espaço, este mapeamento para o espaço de alta dimensão é realizado através de funções *kernel*. Diversos métodos de agrupamento baseados em *kernel* têm sido propostos modificando abordagens já existentes, tais como K -médias, *fuzzy c*-médias, Self Organized Maps – SOM e Neural gas que passaram a incorporar funções *kernel* em suas soluções [13].

Fortalecendo a ideia da utilização das funções *kernel*, diversos estudos demonstraram que os métodos de agrupamento baseados em *kernel*, por produzirem hipersuperfícies não-lineares de separação entre grupos, apresentaram desempenhos melhores do que os métodos convencionais de agrupamento quando a estrutura dos conjuntos de dados é complexa [3, 5, 7, 13].

Nesta perspectiva, [13] apresenta uma variedade de métodos de agrupamento baseados em *kernel*. Estas modificações vêm sendo desenvolvidas sob duas abordagens: kernelização da métrica, em que os centróides dos grupos são obtidos no espaço original das observações de entrada e as distâncias das observações aos centróides dos grupos são calculadas por meio de funções *kernel* e agrupamento no espaço de características, no qual os centróides dos grupos são obtidos no espaço de características.

De acordo com estas abordagens, [26] propôs sob o enfoque de kernelização da métrica, um conjunto de algoritmos do tipo *kernel c*-médias difuso baseado em um *kernel* de Mahalanobis com distâncias quadráticas adaptativas. Os algoritmos baseados no *kernel* gaussiano assumem que os exemplos de conjuntos de dados são mais prováveis de estarem distribuídos dentro de uma região hiperesférica, isto significa que cada variável tem a mesma variância e não há correlação entre variáveis. Contudo, levando em consideração a natureza dos dados que se apresentam nas diversas situações práticas e experimentais é mais provável detectar situações em que os grupos estejam distribuídos dentro de regiões hiperelipsoidais diferentes. Diante deste cenário, a distância de Mahalanobis, que considera as correlações entre variáveis e é invariante em escala, é uma melhor escolha para lidar com dados distribuídos em regiões hiperelipsoidais [8]. Nesta abordagem, as distâncias do tipo Mahalanobis são capazes de lidar com a dispersão dos dados em direções induzidas pelas covariações.

Os algoritmos propostos por [26] foram construídos partindo de quatro distâncias quadráticas adaptativas do tipo Mahalanobis entre observações $\mathbf{x}_i$ e $\mathbf{x}_k$, ambas pertencentes ao $\mathbb{R}^p$. Nesta abordagem, foram utilizadas matrizes de covariâncias completa(não diagonal) e diagonal, única para todos os grupos, como também diferente para cada grupo, respectivamente. Contudo, os métodos propostos se restringem ao caso em que as matrizes de covariâncias são determinadas sob o enfoque da kernelização da métrica.

Nesta dissertação, estendeu-se os métodos apresentados por [26], para o enfoque de agrupamento no espaço de características, baseado no *kernel* de Mahalanobis com distâncias quadráticas adaptativas, isto é, os métodos propostos foram desenvolvidos com matrizes de covariâncias determinadas sob o enfoque agrupamento no espaço de características, onde realizaram um mapeamento de cada observação por meio de uma função não-linear Φ e então obtiveram os centróides dos grupos no espaço de características. A principal vantagem é que, quando mapeado para um espaço de mais alta dimensão, um conjunto de observações de entrada não-linearmente separável torna-se separável linearmente.

Neste cenário, a derivação das expressões matemáticas para a obtenção dos graus de pertinência difuso foi realizada considerando a restrição que os graus de pertinência eram maiores que zero e a soma dos graus de pertinência das variáveis sendo igual a um, enquanto que na derivação das expressões para obtenção das matrizes de covariâncias simétricas definidas positivas $\mathbf{M}$, assumimos o determinante das matrizes sendo igual a um.

Para implementação dos algoritmos, foi utilizada a linguagem de programação **R**. A linguagem **R** é um projeto de código aberto (open source) que está disponível para a maioria das plataformas computacionais. Além de ser uma linguagem de programação, **R** também é um ambiente para computação estatística, modelagem e visualização de dados. Trata-se de uma suíte de softwares integrados que proporcionam facilidades na manipulação de dados, no uso de funções estatísticas e na geração de gráficos [18].

1.1 Objetivos

Objetivos deste trabalho:

- Propor métodos de agrupamento *fuzzy* no espaço de características baseado no *kernel* de Mahalanobis com distâncias quadráticas adaptativas. Nesta abordagem, a distância no espaço de característica foi definida por matrizes de covariâncias completa(não diagonal) e diagonal, única para todo grupo, como também, diferente para cada grupo, respectivamente.
- Dissertar sobre o método *fuzzy c*-médias na perspectiva de agrupamento baseado em *kernel* como embasamento teórico, além de apresentar alguns casos particulares;
- Construir um esquema experimental com conjuntos de dados simulados e reais, avaliando a eficiência dos algoritmos propostos.

1.2 Estrutura do Trabalho

Além deste capítulo introdutório, esta dissertação de mestrado está organizada em mais três capítulos.

- No **Capítulo 2**, apresentamos uma revisão de literatura sobre a teoria básica dos métodos de agrupamentos particionais e agrupamento baseado em *kernel*. Nesta perspectiva é descrito o método *fuzzy c*-médias baseado em kernelização da métrica e agrupamento do espaço de características. Além disso, são apresentados os algoritmos de ambas as abordagens.
- No **Capítulo 3**, introduzimos a principal contribuição deste trabalho: os métodos de agrupamento *fuzzy* no espaço de características, baseados no *kernel* de Mahalanobis, com distâncias quadráticas adaptativas. Para tanto, determinamos a regra de atualização dos centróides no espaço de características, as matrizes de covariâncias simétricas, definidas positivas $\mathbf{M}$, e os graus de pertinência difuso.
- No **Capítulo 4**, apresentamos um desenho de avaliação experimental dos métodos introduzidos no Capítulo 3. Os métodos propostos foram comparados a métodos clássicos de agrupamento como o *fuzzy k*-médias e suas versões baseadas no *Kernel* Gaussiano e os métodos desenvolvido por [26] por meio de experimentos de Monte Carlo, considerando configurações de conjuntos de dados simulados e reais.
- No **Capítulo 5**, finalizamos o trabalho apresentando as conclusões acerca do desempenho dos métodos e as sugestões para trabalhos futuros.

Capítulo 2

Revisão Bibliográfica

Este capítulo está dividido da seguinte forma: na Seção 2.1 é apresentada uma revisão sobre análise e técnica de agrupamento. Na Seção 2.2 são referenciados os métodos particionais, em especial, o método *fuzzy c*-médias difuso baseado em kernelização da métrica e no espaço de características.

2.1 Introdução

A análise desenvolvida com uma perspectiva de separar/organizar observações de acordo com suas características, recebe o nome de análise de agrupamento. A ideia básica dos métodos de agrupamento consiste em reunir em um mesmo grupo observações que apresentam similaridade de acordo com algum critério pré-determinado [27].

Os algoritmos de agrupamento estão divididos em duas técnicas: agrupamento supervisionado que recebe uma influência externa *humana* ou de alguma *regra* que permita conhecer as classes/grupos *a priori*, e agrupamento não-supervisionado. Nesta, a classificação deve ser realizada sem interferências externas.

Estruturando o estudo dos métodos de agrupamento não-supervisionados, destacamos que os mesmos podem ser organizados em dois tipos: algoritmos de agrupamento hierárquicos e algoritmos de agrupamento particionados. Neste trabalho, foram apresentados/discutidos apenas os métodos de agrupamento oriundos das técnicas de agrupamento não-supervisionado do tipo particionado.

2.2 Métodos Particionais

As técnicas de agrupamento do tipo particionais são métodos baseados na minimização de uma função objetivo J , normalmente por meio da otimização (geralmente local) com as observações sendo agrupadas em um número c de grupos escolhidos a

priori. Dessa forma, cada observação é agrupada na classe em que a função objetivo é minimizada.

As técnicas de agrupamento do tipo particionais estão fundamentadas sob dois pilares: agrupamento rígido (*hard*) e agrupamento difuso (*fuzzy*). Nos métodos de agrupamento do tipo rígido, no procedimento de minimização do funcional J , cada observação do banco de dados é associada a um único grupo, isto significa que não se detecta uma interseção entre os grupos. Por outro lado, nos métodos de agrupamento do tipo difuso, uma observação pode pertencer a todos os grupos com um certo grau de pertinência.

Os métodos particionais apresentam uma diferença significativa em relação aos métodos hierárquicos que é a possibilidade de uma observação poder se realocar em qualquer grupo com o desenvolver do algoritmo. Nesta perspectiva, os métodos particionais são extremamente mais rápidos que os métodos hierárquicos [27].

Neste cenário, as medidas de dissimilaridade (ou similaridade) são de extrema importância para a alocação das observações no processo de agrupamento. Um exemplo básico de medidas de dissimilaridade, em métodos particionais, são medidas de distâncias, em que a distância Euclidiana é a mais comumente utilizada.

Uma revisão detalhada dos principais métodos de agrupamento difuso pode ser encontrado em [19]. Nesta perspectiva, o método *fuzzy c-means* (FCM) foi introduzido como uma evolução das técnicas de agrupamento de dados mais recentes e fornece um método que mostra como agrupar observações que pertencem a um espaço multidimensional em um número específico de diferentes agrupamentos [27].

2.2.1 Método *fuzzy c*-médias

O agrupamento *fuzzy* considera um número finito pré determinado de grupos, denotado por c . Aplicando-se o agrupamento *fuzzy* a um conjunto de dados, obtém-se o grau de pertinência de cada elemento em relação a cada um dos c grupos. Nesta visão, o método de agrupamento *fuzzy c*-médias proposto por [4] objetiva estimar estes graus de pertinência.

Um conjunto *fuzzy* Y definido no conjunto não vazio Ω é caracterizado por uma função de pertinência, a qual mapeia os elementos de Ω para o intervalo $[0, 1]$.

$$\begin{aligned} Y : \Omega &\longmapsto [0, 1] \\ n &\longmapsto Y(n). \end{aligned}$$

Desta forma, a função de pertinência associa cada elemento n pertencente à Y um número real $Y(n)$ no intervalo $[0, 1]$, que representa o grau de pertinência do elemento n ao conjunto Y , isto é, o quanto é possível para o elemento n pertencer ao conjunto Y , sendo que, 0 representa a mínima e 1 à máxima pertinência.

O método *fuzzy c*-médias, se inicia com a escolha de c centróides que representam os c grupos a serem formados. Em cada passo do algoritmo, os graus de pertinência são estimados e os centróides são atualizados.

O algoritmo (FCM) objetiva fornecer uma partição *fuzzy* de um conjunto de observações em c grupos $\{\mathbf{P}_1, \dots, \mathbf{P}_c\}$, como também um conjunto de centróides $\{\mathbf{v}_1, \dots, \mathbf{v}_c\}$ de tal forma que um critério J que mede o ajuste entre os grupos e seus centróides são minimizados.

O critério J é definido da seguinte forma,

$$J = \sum_{i=1}^n \sum_{k=1}^c (u_{ki})^m \varphi^2(\mathbf{x}_i, \mathbf{v}_k) = \sum_{i=1}^n \sum_{k=1}^c (u_{ki})^m \sum_{j=1}^p (x_{ij} - v_{kj})^2, \quad (2.1)$$

onde $\varphi^2(\mathbf{x}_i, \mathbf{v}_k)$ é o quadrado da distância euclidiana que mede a dissimilaridade entre um par de vetores, $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ é o vetor de características quantitativas que descreve a i -ésima observação, $\mathbf{v}_k = (v_{k1}, \dots, v_{kp})$ são os centróides dos grupos $\mathbf{P}_k$, com $k = 1, \dots, c$, u_{ki} é o grau de pertinência da i -ésima observação para o k -ésimo grupo $\mathbf{P}_k$ e $m \in (1, \infty)$ é o parâmetro que controla a variação da associação para cada observação $\mathbf{x}_i$, comumente é atribuído o valor 2 a m [27].

Na etapa de representação, a minimização da função objetivo dada pela Eq.(2.1) com respeito a $\mathbf{v}_k$ fornece a seguinte expressão de atualização dos centróide [13, 15]:

$$\mathbf{v}_k = \frac{\sum_{i=1}^n (u_{ki})^m \mathbf{x}_i}{\sum_{i=1}^n (u_{ki})^m}. \quad (2.2)$$

Na etapa de alocação, tomamos uma matriz difusa que contem todos os graus de pertinência difuso da forma $\mathbf{U} = u_{ki}$, onde $k = 1, \dots, c$, e $i = 1, \dots, n$. Para minimizar a Eq.(2.1) introduzimos uma função Lagrangeana com relação à matriz de graus de pertinência $\mathbf{U}$, para cada observação, sob as restrições $u_{ki} > 0$ e $\sum_{k=1}^c u_{ki} = 1$, $\forall k = 1, \dots, c$. Dessa forma, considerando que cada centróide $\mathbf{v}_k$ é fixo a derivada da função lagrangeana

$$L(\mathbf{U}, \alpha) = \sum_{i=1}^n \sum_{k=1}^c (u_{ki})^m \varphi^2(\mathbf{x}_i, \mathbf{v}_k) + \alpha \left(\sum_{k=1}^c u_{ki} - 1 \right), \quad (2.3)$$

calculada com relação a u_{ki} e quando igualada a zero, possibilita obter a seguinte regra de atualização para os graus de pertinência difuso

$$u_{ki} = \left[\sum_{h=1}^c \left(\frac{\sum_{j=1}^p (x_{ij} - v_{kj})^2}{\sum_{l=1}^p (x_{il} - v_{kl})^2} \right)^{\frac{1}{m-1}} \right]^{-1}. \quad (2.4)$$

2.2.2 Agrupamento baseado em *kernel*

Em análise de agrupamento, algoritmos que tem por referência o cálculo da distância Euclidiana entre observações, apresentam um desempenho satisfatório quando direcionados a grupos que são aproximadamente hipersféricos e aproximadamente linearmente separáveis [11].

Em contrapartida, se tivermos uma estrutura de dados complexa, ou seja, grupos com formas que não são hipersféricas e/ou observações não-linearmente separáveis, o desempenho dos métodos pode não ser consistentes. Em virtude desta situação, métodos capazes de trabalhar com dados complexos tem sido propostos, em particular, os métodos de agrupamento baseados em funções *kernel*, tema principal deste trabalho.

Na aprendizagem de máquina, o uso das funções *kernel* foi introduzido por [1] em 1964. A estrutura dos métodos baseados em *kernel* envolve a realização de um mapeamento não linear arbitrário Φ do espaço original p -dimensional $X \subset \mathbb{R}^p$ para um espaço de maior dimensionalidade, considerado espaço de características, $\mathcal{F}$.

O objetivo de realizar o mapeamento em dimensões mais altas é possibilitar a aplicação de um classificador linear no espaço do *kernel*, isto é, poder viabilizar a obtenção de grupos bem definidos e linearmente separáveis. De acordo com o proposto, apresentamos uma breve revisão acerca da teoria básica sobre funções *kernel*.

Kernel de Mercer

Seja $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ um conjunto não-vazio, onde $\mathbf{x}_i \in \mathbb{R}^p$, $\forall i$. Uma função $\mathcal{K} : X \times X \rightarrow \mathbb{R}$ é dita um *kernel* positivo-definido (ou *kernel* de Mercer) se, e somente se, $\mathcal{K}$ é simétrica (i.e. $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_k) = \mathcal{K}(\mathbf{x}_k, \mathbf{x}_i)$) e a seguinte desigualdade é válida [13]:

$$\sum_{i=1}^n \sum_{k=1}^n c_i c_k \mathcal{K}(\mathbf{x}_i, \mathbf{x}_k) \geq 0 \quad \forall n \geq 2, \quad (2.5)$$

em que $c_r \in \mathbb{R} \quad \forall r = 1, \dots, n$.

Dessa forma, para os casos em que temos um conjunto de observações de entrada não-linearmente separável, com um mapeamento não-linear arbitrário para um espaço de características de alta dimensão, podemos torná-lo linearmente separável [17].

Consideremos $\Phi : X \rightarrow \mathcal{F}$ um mapeamento não-linear arbitrário do espaço original X para um espaço de características de alta dimensão $\mathcal{F}$. Nesta situação, aplicando o mapeamento não-linear Φ , o produto interno entre $\mathbf{x}_i$ e $\mathbf{x}_k$ no espaço original pode ser mapeado para o espaço de características por: $\Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_k)$.

A lógica dos métodos baseados em *kernel* é que o mapeamento não-linear Φ não precisa ser explicitamente especificado porque todo *kernel* de Mercer pode ser expresso como

$$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_k) = \Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_k), \quad (2.6)$$

que é usualmente referido como *kernel trick* [25]:

$$\begin{aligned} \|\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_k)\|^2 &= (\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_k))^\top (\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_k)) \\ &= \Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_i) - 2\Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_k) + \Phi(\mathbf{x}_k)^\top \Phi(\mathbf{x}_k) \\ &= \mathcal{K}(\mathbf{x}_i, \mathbf{x}_i) - 2\mathcal{K}(\mathbf{x}_i, \mathbf{x}_k) + \mathcal{K}(\mathbf{x}_k, \mathbf{x}_k). \end{aligned} \quad (2.7)$$

Neste cenário, o cálculo de distâncias de vetores no espaço de características é apenas uma função dos vetores de entrada. Para simplificar a notação, introduzimos a chamada matriz *kernel* $\mathbf{K}$, onde cada elemento pode ser escrito como $\kappa_{ik} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_k)$.

Segue algumas das funções *kernel* tipicamente utilizadas:

- Linear: $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_k) = \mathbf{x}_i^\top \mathbf{x}_k$,
- Polinomial de grau d : $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_k) = (\gamma \mathbf{x}_i^\top \mathbf{x}_k + \theta)^d, \gamma > 0, \theta \geq 0, d \in \mathbb{N}$,
- Gaussiana: $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_k) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_k\|^2}{2\sigma^2}}, \sigma > 0$,

onde γ, σ, θ e d são parâmetros do *kernel*.

Os métodos de agrupamento baseados em *kernel* são estruturados em duas grandes vertentes: kernelização da métrica e agrupamento no espaço de características.

Métodos de agrupamento baseados em kernelização da métrica procuram por centróides no espaço original das observações e a distância entre uma observação $\mathbf{x}_i$ e o centróide do k -ésimo grupo $\mathbf{v}_k$ é calculada por meio de funções *kernel*:

$$\|\Phi(\mathbf{x}_i) - \Phi(\mathbf{v}_k)\|^2 = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_i) - 2\mathcal{K}(\mathbf{x}_i, \mathbf{v}_k) + \mathcal{K}(\mathbf{v}_k, \mathbf{v}_k).$$

Por outro lado, métodos de agrupamento no espaço de características realizam um mapeamento de cada observação por meio de uma função não-linear Φ e então obtêm os centróides dos grupos no espaço de características. Dessa forma, considerando $\mathbf{v}_k^\Phi$ como o centróide do k -ésimo grupo, no espaço de características, alcançamos que é possível calcular $\|\Phi(\mathbf{x}_i) - \mathbf{v}_k^\Phi\|^2$ sem que seja necessário obter $\mathbf{v}_k^\Phi$, através do *kernel trick* Eq.(2.6).

2.2.3 *Kernel* c -médias difuso baseado em kernelização da métrica

O método *kernel* c -médias difuso baseado em kernelização da métrica busca agrupar os dados resolvendo o problema de programação não linear a seguir

$$\begin{aligned} \min_{\mathbf{v}_k, u_{ki}} \quad & J = \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \|\Phi(\mathbf{x}_i) - \Phi(\mathbf{v}_k)\|^2 \\ \text{s.a} \quad & \begin{cases} \sum_{k=1}^c u_{ki} - 1 = 0, \quad \forall i \\ u_{ki} \in [0, 1] \quad \forall k, i \end{cases} \end{aligned} \quad (2.8)$$

em que $\mathbf{v}_k \in \mathbb{R}^p$ é o centróide do k -ésimo grupo, u_{ki} é o grau de pertinência da i -ésima observação para o k -ésimo grupo e $m \in (1, \infty)$ é o parâmetro que controla a variação da associação para cada observação i . Como anteriormente, consideramos $\mathbf{U} = u_{ki}$, matriz de partição difusa.

Temos então que determinar a regra de atualização dos centróides $\mathbf{v}_k$, como também dos graus de pertinência difuso u_{ki} . Inicialmente, a derivação em relação aos centróides dos grupos depende da escolha da função *kernel*. Considerando por exemplo o *kernel* Gaussiano, temos que $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_i) = 1 \quad \forall i$, logo a função objetivo dada pela Eq.(2.8) pode ser reescrita da seguinte forma [16]:

$$J = 2 \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m (1 - \mathcal{K}(\mathbf{x}_i, \mathbf{v}_k)). \quad (2.9)$$

Assim, a regra de atualização dos centróides dos grupos pode ser expressada por:

$$\mathbf{v}_k = \frac{\sum_{i=1}^n (u_{ki})^m \mathcal{K}(\mathbf{x}_i, \mathbf{v}_k) \mathbf{x}_i}{\sum_{i=1}^n (u_{ki})^m \mathcal{K}(\mathbf{x}_i, \mathbf{v}_k)}, \quad \forall k = 1, \dots, c. \quad (2.10)$$

Em seguida, para atualizar a matriz de partição difusa $\mathbf{U}$ mantemos fixos os centróides dos grupos $\mathbf{v}_k$ ($k = 1, \dots, c$). Dessa forma, determinamos os graus de pertinência difuso u_{ki} que minimizam a função objetivo J sob as restrições dadas na Eq.(2.8). Utilizando o método dos multiplicadores de Lagrange, obtemos a seguinte solução [13, 16]:

$$u_{ki} = \left[\sum_{h=1}^c \left(\frac{1 - \mathcal{K}(\mathbf{x}_i, \mathbf{v}_k)}{1 - \mathcal{K}(\mathbf{x}_i, \mathbf{v}_h)} \right)^{\frac{1}{m-1}} \right]^{-1}. \quad (2.11)$$

Com a obtenção da regra de atualização dos centróides e dos graus de pertinência, podemos construir o algoritmo *kernel* c -médias difuso baseado na kernelização da métrica. A execução do algoritmo ocorre de acordo com os seguintes passos:

1. Inicialização

Fixe K (o número de grupos), $2 \leq K < n$; fixe m , $1 < m < \infty$; fixe T (o número máximo de iterações); inicialize os graus de pertinência difusos u_{ki} de tal forma que $u_{ki} \geq 0 \forall k, i$ e $\sum_{k=1}^c u_{ki} = 1, \forall i$; faça $t = 1$.

2. **Etapa 1:** Definição dos melhores centróides dos grupos: Atualize os centróides dos grupos $\mathbf{v}_k$ ($k = 1, \dots, c$) de acordo com a Eq.(2.10).

3. **Etapa 2:** Definição da melhor partição difusa: Atualize os graus de pertinência difusos u_{ki} ($k = 1, \dots, c, i = 1, \dots, n$) de acordo com a Eq.(2.11).

4. Critério de parada

Se $|J^{t+1} - J^t| \leq \varepsilon$ com $\varepsilon > 0$ ou $t > T$, PARE, caso contrário, faça $t = t + 1$ e volte ao passo 2.

2.2.4 Kernel c-médias difuso no espaço de características

O método *kernel c-médias difuso* no espaço de características busca agrupar os dados resolvendo o problema de programação não linear a seguir

$$\begin{aligned} \min_{\mathbf{v}_k^\Phi, u_{ki}} \quad & J = \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \|\Phi(\mathbf{x}_i) - \mathbf{v}_k^\Phi\|^2 \\ \text{s.a} \quad & \begin{cases} \sum_{k=1}^c u_{ki} - 1 = 0, \quad \forall i \\ u_{ki} \in [0, 1] \quad \forall k, i \end{cases} \end{aligned} \quad (2.12)$$

em que u_{ki} é o grau de pertinência da i -ésima observação para o k -ésimo grupo, m é o parâmetro que controla a variação da associação para cada observação e $\mathbf{v}_k^\Phi$ é o centróide do k -grupo no espaço de características.

A minimização da função objetivo dada pela Eq.(2.12) com respeito a $\mathbf{v}_k^\Phi$ fornece a seguinte expressão de atualização dos centróides dos grupos no espaço de características:

$$\mathbf{v}_k^\Phi = \frac{\sum_{i=1}^n (u_{ki})^m \Phi(\mathbf{x}_i)}{\sum_{i=1}^n (u_{ki})^m}, \quad k = 1, \dots, c. \quad (2.13)$$

Em seguida, devemos minimizar a Eq.(2.12) com respeito aos graus de pertinência difuso u_{ki} sob as restrições dadas. Novamente, a utilização do método dos multiplicadores de Lagrange permite determinar a seguinte regra de atualização:

$$u_{ki} = \left[\sum_{h=1}^c \left(\frac{\|\Phi(\mathbf{x}_i) - \mathbf{v}_k^\Phi\|^2}{\|\Phi(\mathbf{x}_i) - \mathbf{v}_h^\Phi\|^2} \right)^{\frac{1}{m-1}} \right]^{-1}. \quad (2.14)$$

Neste caso, a distância $\|\Phi(\mathbf{x}_i) - \mathbf{v}_k^\Phi\|^2$ no espaço de características é calculada através de funções *kernel* no espaço original dos observações, da seguinte forma [13]:

$$\begin{aligned}
\|\Phi(\mathbf{x}_i) - \mathbf{v}_k^\Phi\|^2 &= \Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_i) - 2\Phi(\mathbf{x}_i)^\top (\mathbf{v}_k^\Phi) + (\mathbf{v}_k^\Phi)^\top (\mathbf{v}_k^\Phi) \\
&= \Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_i) - \frac{2 \sum_{l=1}^n (u_{kl})^m \Phi(\mathbf{x}_l)^\top \Phi(\mathbf{x}_i)}{\sum_{l=1}^n (u_{kl})^m} \\
&\quad + \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{kr})^m (u_{ks})^m \Phi(\mathbf{x}_r)^\top \Phi(\mathbf{x}_s)}{(\sum_{r=1}^n (u_{kr})^m)^2} \\
&= \mathcal{K}(\mathbf{x}_i, \mathbf{x}_i) - \frac{2 \sum_{l=1}^n (u_{kl})^m \mathcal{K}(\mathbf{x}_l, \mathbf{x}_i)}{\sum_{l=1}^n (u_{kl})^m} \\
&\quad + \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{kr})^m (u_{ks})^m \mathcal{K}(\mathbf{x}_r, \mathbf{x}_s)}{(\sum_{r=1}^n (u_{kr})^m)^2}. \tag{2.15}
\end{aligned}$$

Adicionalmente, a função objetivo J , dada pela Eq. (2.12), pode ser reescrita da seguinte forma:

$$\begin{aligned}
J = \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \left\{ \mathcal{K}(\mathbf{x}_i, \mathbf{x}_i) - \frac{2 \sum_{l=1}^n (u_{kl})^m \mathcal{K}(\mathbf{x}_l, \mathbf{x}_i)}{\sum_{l=1}^n (u_{kl})^m} \right. \\
\left. + \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{kr})^m (u_{ks})^m \mathcal{K}(\mathbf{x}_r, \mathbf{x}_s)}{(\sum_{r=1}^n (u_{kr})^m)^2} \right\}. \tag{2.16}
\end{aligned}$$

O algoritmo *kernel* c -média difuso no espaço de características não possui a etapa em que os centróides dos grupos são atualizados. A atualização da matriz de partição difusa $\mathbf{U}$ pode ser feita sem o cálculo dos centróides devido ao mapeamento implícito realizado através da função *kernel*. O algoritmo *kernel* c -média difuso no espaço de características é executado de acordo com os seguintes passos:

1. Inicialização

Fixe K (o número de grupos), $2 \leq K < n$; fixe m , $1 < m < \infty$; fixe T (o número máximo de iterações); inicialize os graus de pertinência difusos u_{ki} ($k = 1, \dots, c, i = 1, \dots, n$) de tal forma que $u_{ki} \geq 0 \forall k, i$ e $\sum_{k=1}^c u_{ki} = 1, \forall i$; escolha uma função *kernel* adequada $\mathcal{K}$, calcule a matriz *kernel* $\mathbf{K}$ e faça $t = 1$.

2. **Etapa 1:** Definição da melhor partição difusa. Atualize os graus de pertinência difusos u_{ki} de acordo com a Eq.(2.14).

3. Critério de parada

Se $|J^{t+1} - J^t| \leq \varepsilon$ com $\varepsilon > 0$ ou $t > T$, PARE, caso contrário, faça $t = t + 1$ e volte ao passo 2.

As propriedades de convergências deste tipo de algoritmo podem ser demonstradas como em [10].

Capítulo 3

Métodos Propostos

Este capítulo está dividido em duas seções: na Seção 3.1 são apresentadas algumas considerações sobre distância de Mahalanobis e *kernel* de Mahalanobis. Na seção Seção 3.2 que é a principal, são propostos os métodos de agrupamento no espaço de características baseados no *kernel* de Mahalanobis, com distâncias quadráticas adaptativas. Para tanto, determinamos a regra de atualização dos centróides no espaço de característica $\mathbf{v}_k^\Phi$, as matrizes de covariâncias completa e diagonal $\mathbf{M}$ e os graus de pertinência difuso u_{ki} .

3.1 Considerações Iniciais quanto a Distância de Mahalanobis

A distância de Mahalanobis é uma métrica que leva em consideração a correlação entre os conjuntos de dados, no qual distintos padrões podem ser identificados e analisados [22]. Nesta perspectiva, a distância de Mahalanobis entre uma observação $\mathbf{x}_i$ e o centróide $\mathbf{v}$, ambos pertencentes ao $\mathbb{R}^p$, tem a seguinte representação

$$d_{\Sigma^{-1}}^2(\mathbf{x}_i, \mathbf{v}) = (\mathbf{x}_i, \mathbf{v})^\top \Sigma^{-1} (\mathbf{x}_i, \mathbf{v}), \quad (3.1)$$

sendo que

$$\mathbf{v} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \quad \text{e} \quad \Sigma = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i, \mathbf{v})(\mathbf{x}_i, \mathbf{v})^\top.$$

De acordo com a definição de distância de Mahalanobis, podemos definir um *kernel* adaptativo de Mahalanobis e dessa forma substituir a distância euclidiana no

kernel gaussiano por uma do tipo-Mahalanobis [8]:

$$\mathcal{K}^{(M)}(\mathbf{x}_i, \mathbf{x}_k) = \exp \left\{ -\frac{(\mathbf{x}_i - \mathbf{x}_k)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_k)}{2\sigma^2} \right\}, \quad (3.2)$$

em que $d_M^2(\mathbf{x}_i, \mathbf{x}_k) = (\mathbf{x}_i - \mathbf{x}_k)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_k)$ é uma distância quadrática definida por uma matriz simétrica definida positiva $\mathbf{M}$. Neste caso, a distância de Mahalanobis é verificada entre duas observações $\mathbf{x}_i$ e $\mathbf{x}_k$ no $\mathbb{R}^p$. Podemos observar ainda que, para a situação em que a matriz de covariâncias $\mathbf{M} = \mathbf{I}$, em que $\mathbf{I}$ caracteriza a matriz identidade de ordem $p \times p$, obtemos a expressão relativa ao *kernel* Gaussiano, e dessa forma, podemos dizer que o *kernel* de Mahalanobis é uma extensão do *kernel* Gaussiano. Além disso, a distância de Mahalanobis pode ser compreendida como invariante com relação a escala, ou seja, a mesma distância é obtida independentemente da unidade utilizada para caracterizar a posição das variáveis.

Em referência ao estudo da distância de Mahalanobis, abordamos quatro distâncias quadráticas adaptativas do tipo-Mahalanobis entre observações $\mathbf{x}_i$ e $\mathbf{x}_k$, ambas pertencentes ao $\mathbb{R}^p$. Nesta abordagem, foram utilizadas matrizes de covariâncias completas e diagonal única, isto é, comum a todos os grupos, como também, diferente para cada grupo, respectivamente. Dessa forma, temos

- (a) Distância quadrática adaptativa definida por uma matriz de covariâncias completa única para todos os grupos:

$$\mathbf{M}_k = \mathbf{M} : d_M^2(\mathbf{x}_i, \mathbf{x}_k) = (\mathbf{x}_i - \mathbf{x}_k)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_k), k = 1, \dots, c. \quad (3.3)$$

- (b) Distância quadrática adaptativa definida por uma matriz de covariâncias diagonal única para todos os grupos:

$$\begin{aligned} \mathbf{M}_k = \mathbf{M} &= \text{diag} \{ \lambda_1, \dots, \lambda_p \} : d_M^2(\mathbf{x}_i, \mathbf{x}_k) = (\mathbf{x}_i - \mathbf{x}_k)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_k) \\ &= \sum_{j=1}^p \lambda_j \varphi^2(\mathbf{x}_{ij}, \mathbf{x}_{kj}), k = 1, \dots, c. \end{aligned} \quad (3.4)$$

- (c) Distância quadrática adaptativa definida por uma matriz de covariâncias completa para cada grupo:

$$\mathbf{M}_k : d_{\mathbf{M}_k}^2(\mathbf{x}_i, \mathbf{x}_k) = (\mathbf{x}_i - \mathbf{x}_k)^\top \mathbf{M}_k(\mathbf{x}_i - \mathbf{x}_k), k = 1, \dots, c. \quad (3.5)$$

- (d) Distância quadrática adaptativa definida por uma matriz de covariâncias dia-

gonal para cada grupo:

$$\begin{aligned} \mathbf{M}_k &= \text{diag} \{ \lambda_{k1}, \dots, \lambda_{kp} \} : d_{\mathbf{M}_k}^2(\mathbf{x}_i, \mathbf{x}_k) = (\mathbf{x}_i - \mathbf{x}_k)^\top \mathbf{M}_k (\mathbf{x}_i - \mathbf{x}_k) \\ &= \sum_{j=1}^p \lambda_{kj} \varphi^2(\mathbf{x}_{ij}, \mathbf{x}_{kj}), k = 1, \dots, c. \end{aligned} \quad (3.6)$$

A partir da mudança das matrizes de covariâncias apresentadas em (3.3), (3.4), (3.5) e (3.6) e da definição do *Kernel* adaptativo de Mahalanobis, introduzimos os métodos de agrupamento *fuzzy* no espaço de características baseado no *Kernel* de Mahalanobis com distâncias quadráticas adaptativas.

Dada uma distância quadrática adaptativa, que pode ser única ou diferente para cada grupo, os métodos de agrupamento *fuzzy* procuram uma partição deste tipo em um conjunto de observações. São formados c grupos, onde cada representante de grupo é um centróide, portanto teremos um conjunto de c centróides $\{\mathbf{v}_1, \dots, \mathbf{v}_c\}$. O critério para a formação dos grupos é que a distância entre os elementos do grupo e seu centróide seja minimizada.

3.2 Métodos de agrupamento *fuzzy* no espaço de características baseado no *kernel* de Mahalanobis com distâncias quadráticas adaptativas

Os algoritmos introduzidos nesta seção otimizam uma função objetivo não linear J , que mede o ajuste entre os grupos e seus centróides, como pode ser visto no modelo abaixo.

$$\begin{aligned} \min_{\mathbf{v}_k^\Phi, \mathbf{M}, u_{ki}} \quad & J = \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \phi^2(\mathbf{x}_i, \mathbf{v}_k^\Phi), \\ \text{s.a} \quad & \begin{cases} \sum_{k=1}^c u_{ki} - 1 = 0, \quad \forall k, i \\ u_{ki} \in [0, 1] \\ \det(\mathbf{M}) - 1 = 0, \quad \forall \mathbf{M} \end{cases} \end{aligned} \quad (3.7)$$

em que $\phi^2(\mathbf{x}_i, \mathbf{v}_k^\Phi)$ se apresenta como uma medida de distância adequada entre um padrão $\mathbf{x}_i$ e o centróide do k -ésimo grupo $\mathbf{v}_k^\Phi$ no espaço de características, determinada por meio de uma função *kernel* de Mahalanobis, sendo que u_{ki} , $k = 1, \dots, c$, e $i = 1, \dots, n$, e m são definidos como na Seção 2.2.4. Considerando as medidas de distâncias propostas na Seção 3.1, temos:

- (a) Método de agrupamento *fuzzy*, no espaço de características, baseado no *kernel* de Mahalanobis com distância quadrática adaptativa definida por uma matriz de covariâncias *fuzzy* completa (KFCM.FS.GF):

$$J_1 = \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \|\Phi(\mathbf{x}_i) - \mathbf{v}_k^\Phi\|^2$$

$$J_1 = \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \left\{ 1 - \frac{2 \sum_{l=1}^n (u_{kl})^m \mathcal{K}^{(M)}(\mathbf{x}_l, \mathbf{x}_i)}{\sum_{l=1}^n (u_{kl})^m} + \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{kr})^m (u_{ks})^m \mathcal{K}^{(M)}(\mathbf{x}_r, \mathbf{x}_s)}{(\sum_{r=1}^n (u_{kr})^m)^2} \right\}. \quad (3.8)$$

Nesta perspectiva, temos que

$$\mathcal{K}^{(M)}(\mathbf{x}_i, \mathbf{x}_k) = \exp \left\{ -\frac{(\mathbf{x}_i - \mathbf{x}_k)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_k)}{2\sigma^2} \right\},$$

com

$$d_M^2(\mathbf{x}_i, \mathbf{x}_k) = (\mathbf{x}_i - \mathbf{x}_k)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_k),$$

sendo uma distância quadrática definida por uma matriz simétrica definida positiva $\mathbf{M}$, que é a mesma para todos os grupos.

- (b) Método de agrupamento *fuzzy*, no espaço de características, baseado no *kernel* de Mahalanobis com distância quadrática adaptativa definida por uma matriz de covariâncias *fuzzy* diagonal (KFCM.FS.GD):

$$J_2 = \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \|\Phi(\mathbf{x}_i) - \mathbf{v}_k^\Phi\|^2$$

$$J_2 = \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \left\{ 1 - \frac{2 \sum_{l=1}^n (u_{kl})^m \mathcal{K}^{(M)}(\mathbf{x}_l, \mathbf{x}_i)}{\sum_{l=1}^n (u_{kl})^m} + \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{kr})^m (u_{ks})^m \mathcal{K}^{(M)}(\mathbf{x}_r, \mathbf{x}_s)}{(\sum_{r=1}^n (u_{kr})^m)^2} \right\}, \quad (3.9)$$

onde $\mathcal{K}^{(M)}(\mathbf{x}_i, \mathbf{x}_k)$ é definido pela Eq.(3.2) e a matriz $\mathbf{M}$ sendo definida da seguinte forma:

$$\mathbf{M} = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_p \end{pmatrix},$$

com isso, temos uma modificação no método em que $d_M^2(\mathbf{x}_i, \mathbf{x}_k)$ é uma distância quadrática definida por uma matriz simétrica diagonal definida positiva $\mathbf{M}$, que é a mesma para todos os grupos.

- (c) Método de agrupamento *fuzzy* no espaço de características baseado no *kernel* de Mahalanobis com distâncias quadráticas adaptativas diferentes para cada grupo definidas por uma matriz de covariâncias *fuzzy* completa (KFCM.FS.LF):

$$J_3 = \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \|\Phi(\mathbf{x}_i) - \mathbf{v}_k^\Phi\|^2$$

$$J_3 = \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \left\{ 1 - \frac{2 \sum_{l=1}^n (u_{kl})^m \mathcal{K}^{(M)}(\mathbf{x}_l, \mathbf{x}_i)}{\sum_{l=1}^n (u_{kl})^m} + \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{kr})^m (u_{ks})^m \mathcal{K}^{(M)}(\mathbf{x}_r, \mathbf{x}_s)}{(\sum_{r=1}^n (u_{kr})^m)^2} \right\}, \quad (3.10)$$

sendo que,

$$\mathcal{K}^{(M)}(\mathbf{x}_i, \mathbf{x}_k) = \exp \left\{ -\frac{(\mathbf{x}_i - \mathbf{x}_k)^\top \mathbf{M}_k (\mathbf{x}_i - \mathbf{x}_k)}{2\sigma^2} \right\},$$

com

$$d_{M_k}^2(\mathbf{x}_i, \mathbf{x}_k) = (\mathbf{x}_i - \mathbf{x}_k)^\top \mathbf{M}_k (\mathbf{x}_i - \mathbf{x}_k),$$

sendo uma distância quadrática definida por uma matriz simétrica definida positiva $\mathbf{M}_k$, diferente para cada grupo.

- (d) Método de agrupamento *fuzzy* no espaço de características baseado no *kernel* de Mahalanobis, com distâncias quadráticas adaptativas diferentes para cada grupo, definidas por uma matriz de covariâncias *fuzzy* diagonal (KFCM.FS.LD):

$$J_4 = \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \|\Phi(\mathbf{x}_i) - \mathbf{v}_k^\Phi\|^2$$

$$J_4 = \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \left\{ 1 - \frac{2 \sum_{l=1}^n (u_{kl})^m \mathcal{K}^{(M)}(\mathbf{x}_l, \mathbf{x}_i)}{\sum_{l=1}^n (u_{kl})^m} + \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{kr})^m (u_{ks})^m \mathcal{K}^{(M)}(\mathbf{x}_r, \mathbf{x}_s)}{(\sum_{r=1}^n (u_{kr})^m)^2} \right\}, \quad (3.11)$$

sendo que,

$$\mathcal{K}^{(M)}(\mathbf{x}_i, \mathbf{x}_k) = \exp \left\{ -\frac{(\mathbf{x}_i - \mathbf{x}_k)^\top \mathbf{M}_k (\mathbf{x}_i - \mathbf{x}_k)}{2\sigma^2} \right\},$$

e a matriz $\mathbf{M}_k$ é definida da seguinte forma:

$$\mathbf{M}_k = \begin{pmatrix} \lambda_{k1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_{kp} \end{pmatrix},$$

com isso, $d_{M_k}^2(\mathbf{x}_i, \mathbf{x}_k)$ é uma distância quadrática definida por uma matriz simétrica diagonal definida positiva $\mathbf{M}_k$, que é diferente para cada grupo.

Com a apresentação dos métodos de agrupamento *fuzzy* baseados no *Kernel* de Mahalanobis, organizamos o algoritmo em três etapas. A primeira consistiu em determinar a regra de atualização dos centróides no espaço de características. Nesta etapa, derivamos o funcional J dado por (3.7) em relação ao centróide do k -ésimo grupo $\mathbf{v}_k^\Phi$ no espaço de características, neste processo os graus de pertinência difuso u_{ki} , com $k = 1, \dots, c$ e $i = 1, \dots, n$, e as matrizes de covariâncias *fuzzy* são mantidos fixos.

Proposição 3.2.1. *Qualquer que seja a medida de distância adotada Eqs. (3.3), (3.4), (3.5), (3.6), o centróide do k -ésimo grupo no espaço de características $\mathbf{v}_k^\Phi$ ($k = 1, \dots, c$), que minimiza o critério J , dado pela Eq. (3.7) é atualizado de acordo com a seguinte equação:*

$$\mathbf{v}_k^\Phi = \frac{\sum_{i=1}^n (u_{ki})^m \Phi(\mathbf{x}_i)}{\sum_{i=1}^n (u_{ki})^m}, \quad (k = 1, \dots, c). \quad (3.12)$$

Demonstração. Para o k -ésimo grupo o problema se torna encontrar o componente $\mathbf{v}_k^\Phi$ que maximiza o termo

$$\sum_{i=1}^n (u_{ki})^m \|\Phi(\mathbf{x}_i) - \mathbf{v}_k^\Phi\|^2. \quad (3.13)$$

Dessa forma, tomando a primeira derivada do termo dado em (3.13) com relação a $\mathbf{v}_k^\Phi$ igual a zero, temos, para ($k = 1, \dots, c$),

$$\begin{aligned} \frac{\partial J}{\partial \mathbf{v}_k^\Phi} &= \sum_{i=1}^n (u_{ki})^m \frac{\partial \|\Phi(\mathbf{x}_i) - \mathbf{v}_k^\Phi\|^2}{\partial \mathbf{v}_k^\Phi} = 0 \\ -2 \sum_{i=1}^n (u_{ki})^m (\Phi(\mathbf{x}_i) - \mathbf{v}_k^\Phi) &= 0. \end{aligned}$$

Resolvendo, tem-se

$$\mathbf{v}_k^\Phi = \frac{\sum_{i=1}^n (u_{ki})^m \Phi(\mathbf{x}_i)}{\sum_{i=1}^n (u_{ki})^m}.$$

□

Na segunda etapa do algoritmo, determinamos as matrizes de covariâncias $\mathbf{M}$ correspondente a cada distância adaptativa proposta nesta dissertação. Para isso, foram identificadas as expressões que minimizam os seguintes critérios de agrupamentos das Eqs.((3.8),(3.9),(3.10),(3.11)).

Proposição 3.2.2. *As matrizes simétricas definitivas positivas $\mathbf{M}$, que minimizam a função objetivo J dada pela Eq.(3.7) com a restrição de $\det(\mathbf{M}) = 1$, são atualizadas de acordo com as seguintes expressões:*

- (a) Matriz de covariâncias *fuzzy* completa única que minimiza o critério J dado pela Eq.(3.8):

$$\mathbf{M} = [\det(\mathbf{D})]^{\frac{1}{p}} (\mathbf{D})^{-1}, \quad (3.14)$$

em que,

$$\mathbf{D} = \sum_{k=1}^c \mathbf{E}_k$$

e

$$\mathbf{E}_k = \sum_{i=1}^n (u_{ki})^m \left(\frac{\sum_{l=1}^n (u_{kl})^m \mathcal{K}^M(\mathbf{x}_l, \mathbf{x}_i) (\mathbf{x}_l - \mathbf{x}_i) (\mathbf{x}_l - \mathbf{x}_i)^\top}{\sum_{l=1}^n (u_{kl})^m} - \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{kr})^m (u_{ks})^m \mathcal{K}^M(\mathbf{x}_r, \mathbf{x}_s) (\mathbf{x}_r - \mathbf{x}_s) (\mathbf{x}_r - \mathbf{x}_s)^\top}{2(\sum_{r=1}^n (u_{kr})^m)^2} \right). \quad (3.15)$$

- (b) Matriz diagonal de covariâncias *fuzzy* única que minimiza o critério J dado pela Eq.(3.9):

$$\mathbf{M} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{pmatrix} = [\det(\mathbf{D})]^{\frac{1}{p}} \mathbf{D}^{-1}, \quad (3.16)$$

em que,

$$\mathbf{D} = \sum_{k=1}^c \mathbf{E}_k$$

e

$$\begin{aligned} \mathbf{E}_k &= \text{diag} \left(\sum_{i=1}^n (u_{ki})^m \left(\frac{\sum_{l=1}^n (u_{kl})^m \mathcal{K}^M(\mathbf{x}_l, \mathbf{x}_i)(\mathbf{x}_l - \mathbf{x}_i)(\mathbf{x}_l - \mathbf{x}_i)^\top}{\sum_{l=1}^n (u_{kl})^m} \right. \right. \\ &\quad \left. \left. - \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{kr})^m (u_{ks})^m \mathcal{K}^M(\mathbf{x}_r, \mathbf{x}_s)(\mathbf{x}_r - \mathbf{x}_s)(\mathbf{x}_r - \mathbf{x}_s)^\top}{2(\sum_{r=1}^n (u_{kr})^m)^2} \right) \right). \end{aligned}$$

(c) Matriz de covariâncias *fuzzy* completa diferente para cada grupo que minimiza o critério J dado pela Eq.(3.10):

$$\mathbf{M}_k = [\det(\mathbf{D}_k)]^{\frac{1}{p}} \mathbf{D}_k^{-1}, \quad (3.17)$$

em que,

$$\mathbf{D}_k = \sum_{k=1}^c \mathbf{E}_k$$

e

$$\begin{aligned} \mathbf{E}_k &= \sum_{i=1}^n (u_{ki})^m \left(\frac{\sum_{l=1}^n (u_{kl})^m \mathcal{K}^M(\mathbf{x}_l, \mathbf{x}_i)(\mathbf{x}_l - \mathbf{x}_i)(\mathbf{x}_l - \mathbf{x}_i)^\top}{\sum_{l=1}^n (u_{kl})^m} \right. \\ &\quad \left. - \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{kr})^m (u_{ks})^m \mathcal{K}^M(\mathbf{x}_r, \mathbf{x}_s)(\mathbf{x}_r - \mathbf{x}_s)(\mathbf{x}_r - \mathbf{x}_s)^\top}{2(\sum_{r=1}^n (u_{kr})^m)^2} \right). \end{aligned}$$

(d) Matriz diagonal de covariâncias *fuzzy* diferente para cada grupo que minimiza o critério J dado pela Eq.(3.11):

$$\mathbf{M}_k = \begin{pmatrix} \lambda_{k1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_{kp} \end{pmatrix} = [\det(\mathbf{D}_k)]^{\frac{1}{p}} \mathbf{D}_k^{-1}, \quad (3.18)$$

sendo que

$$\mathbf{D}_k = \sum_{k=1}^c \mathbf{E}_k$$

e

$$\begin{aligned} \mathbf{E}_k &= \text{diag} \left(\sum_{i=1}^n (u_{ki})^m \left(\frac{\sum_{l=1}^n (u_{kl})^m \mathcal{K}^M(\mathbf{x}_l, \mathbf{x}_i)(\mathbf{x}_l - \mathbf{x}_i)(\mathbf{x}_l - \mathbf{x}_i)^\top}{\sum_{l=1}^n (u_{kl})^m} \right. \right. \\ &\quad \left. \left. - \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{kr})^m (u_{ks})^m \mathcal{K}^M(\mathbf{x}_r, \mathbf{x}_s)(\mathbf{x}_r - \mathbf{x}_s)(\mathbf{x}_r - \mathbf{x}_s)^\top}{2(\sum_{r=1}^n (u_{kr})^m)^2} \right) \right). \end{aligned}$$

Demonstração. (a) Para este primeiro caso, os graus de pertinência difusos u_{ki} ($k = 1, \dots, c, i = 1, \dots, n$), o parâmetro m e os centróides $\mathbf{v}_k^\Phi$ dos grupos no espaço de características são mantidos fixos, podemos então reescrever o critério J_1 como uma

função em relação a $\mathbf{M}$,

$$\sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \left\{ 1 - \frac{2 \sum_{l=1}^n (u_{kl})^m \mathcal{K}^M(\mathbf{x}_l, \mathbf{x}_i)}{\sum_{l=1}^n (u_{kl})^m} + \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{kr})^m (u_{ks})^m \mathcal{K}^M(\mathbf{x}_r, \mathbf{x}_s)}{(\sum_{r=1}^n (u_{kr})^m)^2} \right\},$$

considerando a distância de Mahalanobis, temos a seguinte representação para J_1

$$\begin{aligned} \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m & \left\{ 1 - \frac{2 \sum_{l=1}^n (u_{kl})^m \exp \left\{ -\frac{(\mathbf{x}_l - \mathbf{x}_i)^\top \mathbf{M} (\mathbf{x}_l - \mathbf{x}_i)}{2\sigma^2} \right\}}{\sum_{l=1}^n (u_{kl})^m} \right. \\ & \left. + \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{kr})^m (u_{ks})^m \exp \left\{ -\frac{(\mathbf{x}_r - \mathbf{x}_s)^\top \mathbf{M} (\mathbf{x}_r - \mathbf{x}_s)}{2\sigma^2} \right\}}{(\sum_{r=1}^n (u_{kr})^m)^2} \right\}, \end{aligned}$$

para organizar melhor o processo de minimização, fizemos a seguinte mudança dos termos

$$\begin{aligned} A &= \frac{2 \sum_{l=1}^n (u_{kl})^m \exp \left\{ -\frac{(\mathbf{x}_l - \mathbf{x}_i)^\top \mathbf{M} (\mathbf{x}_l - \mathbf{x}_i)}{2\sigma^2} \right\}}{\sum_{l=1}^n (u_{kl})^m} \\ B &= \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{kr})^m (u_{ks})^m \exp \left\{ -\frac{(\mathbf{x}_r - \mathbf{x}_s)^\top \mathbf{M} (\mathbf{x}_r - \mathbf{x}_s)}{2\sigma^2} \right\}}{(\sum_{r=1}^n (u_{kr})^m)^2} \end{aligned}$$

neste cenário, alcançamos que o funcional J_1 pode ser reescrito como

$$J = \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \{1 - A + B\}.$$

Além disso, fazendo, $g(\mathbf{M}) = 1 - \det(\mathbf{M})$, determinamos os extremos de $J_1(\mathbf{M})$ com a restrição $g(\mathbf{M}) = 0$. Para isso, consideramos a função Lagrangeana

$$L(\mathbf{M}, \alpha) = J(\mathbf{M}) + \alpha g(\mathbf{M}),$$

considerando $C = \alpha g(\mathbf{M}) = \alpha(1 - \det(\mathbf{M}))$, obtemos o funcional

$$L(\mathbf{M}, \alpha) = \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \{1 - A + B\} + C, \quad (3.19)$$

derivando os termos da Eq.(3.19) em relação a $\mathbf{M}$, obtemos

$$\frac{\partial L(\mathbf{M}, \alpha)}{\partial \mathbf{M}} = \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \left\{ -\frac{\partial A}{\partial \mathbf{M}} + \frac{\partial B}{\partial \mathbf{M}} \right\} + \frac{\partial C}{\partial \mathbf{M}}$$

onde

$$\begin{aligned}\frac{\partial A}{\partial \mathbf{M}} &= -\frac{1}{\sigma^2} \frac{\sum_{l=1}^n (u_{kl})^m \mathcal{K}^M(\mathbf{x}_l, \mathbf{x}_i)(\mathbf{x}_l - \mathbf{x}_i)(\mathbf{x}_l - \mathbf{x}_i)^\top}{\sum_{l=1}^n (u_{kl})^m} \\ \frac{\partial B}{\partial \mathbf{M}} &= -\frac{1}{\sigma^2} \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{kr})^m (u_{ks})^m \mathcal{K}^M(\mathbf{x}_r, \mathbf{x}_s)(\mathbf{x}_r - \mathbf{x}_s)(\mathbf{x}_r - \mathbf{x}_s)^\top}{2(\sum_{r=1}^n (u_{kr})^m)^2} \\ \frac{\partial C}{\partial \mathbf{M}} &= -\alpha \det(\mathbf{M}) \mathbf{M}^{-1}\end{aligned}$$

fazendo $\frac{\partial L(\mathbf{M}, \alpha)}{\partial \mathbf{M}} = 0$, chegamos na equação a seguir:

$$\begin{aligned}-\frac{\partial C}{\partial \mathbf{M}} &= \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \left\{ -\frac{\partial A}{\partial \mathbf{M}} + \frac{\partial B}{\partial \mathbf{M}} \right\} \\ \alpha \det(\mathbf{M}) \mathbf{M}^{-1} &= \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \left\{ -\frac{\partial A}{\partial \mathbf{M}} + \frac{\partial B}{\partial \mathbf{M}} \right\}\end{aligned}$$

minimizando sob a seguinte restrição $\det(\mathbf{M}) = 1$, temos então

$$\begin{aligned}\mathbf{M}^{-1} &= \frac{1}{\alpha} \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \left\{ -\frac{\partial A}{\partial \mathbf{M}} + \frac{\partial B}{\partial \mathbf{M}} \right\} \\ \mathbf{M}^{-1} &= \frac{1}{\alpha} \sum_{k=1}^c \left(\sum_{i=1}^n (u_{ki})^m \left(-\frac{\partial A}{\partial \mathbf{M}} + \frac{\partial B}{\partial \mathbf{M}} \right) \right), \\ \mathbf{M}^{-1} &= \frac{1}{\alpha \sigma^2} \sum_{k=1}^c \sum_{i=1}^n (u_{ki})^m \left(\frac{\sum_{l=1}^n (u_{kl})^m \mathcal{K}^M(\mathbf{x}_l, \mathbf{x}_i)(\mathbf{x}_l - \mathbf{x}_i)(\mathbf{x}_l - \mathbf{x}_i)^\top}{\sum_{l=1}^n (u_{kl})^m} \right. \\ &\quad \left. - \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{kr})^m (u_{ks})^m \mathcal{K}^M(\mathbf{x}_r, \mathbf{x}_s)(\mathbf{x}_r - \mathbf{x}_s)(\mathbf{x}_r - \mathbf{x}_s)^\top}{2(\sum_{r=1}^n (u_{kr})^m)^2} \right),\end{aligned}$$

fazendo

$$\begin{aligned}\mathbf{E}_k &= \sum_{i=1}^n (u_{ki})^m \left(\frac{\sum_{l=1}^n (u_{kl})^m \mathcal{K}^M(\mathbf{x}_l, \mathbf{x}_i)(\mathbf{x}_l - \mathbf{x}_i)(\mathbf{x}_l - \mathbf{x}_i)^\top}{\sum_{l=1}^n (u_{kl})^m} \right. \\ &\quad \left. - \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{kr})^m (u_{ks})^m \mathcal{K}^M(\mathbf{x}_r, \mathbf{x}_s)(\mathbf{x}_r - \mathbf{x}_s)(\mathbf{x}_r - \mathbf{x}_s)^\top}{2(\sum_{r=1}^n (u_{kr})^m)^2} \right)\end{aligned}$$

determinamos

$$\begin{aligned}\mathbf{M}^{-1} &= \left(\frac{1}{\alpha \sigma^2} \right) \sum_{k=1}^c \mathbf{E}_k \\ \mathbf{M}^{-1} &= \left(\frac{1}{\alpha \sigma^2} \right) \mathbf{D}, \text{ onde } \mathbf{D} = \sum_{k=1}^c \mathbf{E}_k.\end{aligned}$$

Sabendo que $\det(\mathbf{M}^{-1}) = \frac{1}{\det(\mathbf{M})} = 1$, podemos fazer,

$$\mathbf{M}^{-1} = \left(\frac{1}{\alpha \sigma^2} \right) \mathbf{D} \implies \det(\mathbf{M}^{-1}) = \left(\frac{1}{\alpha^p \sigma^{2p}} \right) \det(\mathbf{D}) = 1 \implies \alpha = \frac{[\det(\mathbf{D})]^{\frac{1}{p}}}{\sigma^2},$$

dessa forma, obtemos a seguinte matriz,

$$\mathbf{M}^{-1} = \left(\frac{1}{[\det(\mathbf{D})]^{\frac{1}{p}}} \right) \mathbf{D} = \frac{\mathbf{D}}{[\det(\mathbf{D})]^{\frac{1}{p}}},$$

donde nos permite determinar que

$$\mathbf{M} = [\det(\mathbf{D})]^{\frac{1}{p}} \mathbf{D}^{-1}.$$

□

Seguindo o mesmo raciocínio empregado em (a), conseguimos demonstrar as partes (b), (c) e (d). Observamos que os demais métodos a serem demonstrados são casos especiais do método J_1 , em (c) obtemos matrizes de covariâncias completas e diferentes para cada grupo, ao mesmo passo que em (b) e (d) determinamos matrizes diagonais de covariâncias *fuzzy*, única e diferente para cada grupo, respectivamente.

Para finalizar, a terceira etapa do algoritmo, reservamos para determinar os graus de pertinência difuso u_{ki} ($k = 1, \dots, c, i = 1, \dots, n$). Para isso, os centróides dos grupos no espaço de características $\mathbf{v}_k^\Phi$ ($k = 1, \dots, c$), são mantidos fixos. O problema se torna, então, encontrar os graus que minimizam o critério J sob restrições adequadas.

Proposição 3.2.3. *Os graus de pertinência fuzzy u_{ki} ($k = 1, \dots, c, i = 1, \dots, n$), que minimizam a função objetivo J dada pela Eq.(3.7), sujeito a $u_{ki} \in [0, 1] \forall k, i$ e $\sum_{k=1}^c u_{ki} = 1 \forall i$, são atualizados de acordo com a seguinte equação:*

$$u_{ki} = \left[\sum_{h=1}^c \left(\frac{\|\Phi(\mathbf{x}_i) - \mathbf{v}_k^\Phi\|^2}{\|\Phi(\mathbf{x}_i) - \mathbf{v}_h^\Phi\|^2} \right)^{\frac{1}{m-1}} \right]^{-1}. \quad (3.20)$$

Demonstração. De acordo com as condições, temos que $u_{ki} \in [0, 1] \forall k, i$ e $\sum_{k=1}^c u_{ki} = 1 \forall i$. Além disso, sabendo que os centróides no espaço de característica $\mathbf{v}_k^\Phi$ ($k = 1, \dots, c$), o parâmetro m e a matriz $\mathbf{M}$ são fixos, podemos reescrever J da forma

$$J(\mathbf{u}_1, \dots, \mathbf{u}_c) = \sum_{k=1}^c J_k(\mathbf{u}_k), \quad J_k(\mathbf{u}_k) = J_k(u_{k1}, \dots, u_{kn}) = \sum_{i=1}^n (u_{ki})^m J_k,$$

onde $J_k = \|\Phi(\mathbf{x}_i) - \mathbf{v}_k^\Phi\|^2$.

Em virtude do critério J ser aditivo, o problema se torna minimizar J_k , ($k = 1, \dots, c$). Considerando $g_k(u_{k1}, \dots, u_{kn}) = \sum_{i=1}^n u_{ki} - 1$, determinamos os extremos de $J_k(u_{k1}, \dots, u_{kn})$ sujeito a restrição $g_k(u_{k1}, \dots, u_{kn}) = 0$. Para tanto, aplicamos o método dos multiplicadores de Lagrange para resolver o seguinte sistema

$$\nabla J_k(u_{k1}, \dots, u_{kn}) = \mu \nabla g_k(u_{k1}, \dots, u_{kn}).$$

Dessa forma, para $k = 1, \dots, c$ e $i = 1, \dots, n$, temos

$$\begin{aligned} \frac{\partial J_k(u_{k1}, \dots, u_{kn})}{\partial u_{ki}} &= \mu \frac{\partial g_k(u_{k1}, \dots, u_{kn})}{\partial u_{ki}} \\ m(u_{ki})^{m-1} J_k &= \mu \\ u_{ki} &= \left(\frac{\mu}{m}\right)^{\frac{1}{m-1}} \cdot \frac{1}{(J_k)^{\frac{1}{m-1}}}. \end{aligned} \quad (3.21)$$

Sabendo que $\sum_{h=1}^c u_{hi} = 1$, obtemos

$$\begin{aligned} \sum_{h=1}^c \left(\frac{\mu}{m}\right)^{\frac{1}{m-1}} \cdot \frac{1}{(J_h)^{\frac{1}{m-1}}} &= 1 \\ \left(\frac{\mu}{m}\right)^{\frac{1}{m-1}} &= \frac{1}{\sum_{h=1}^c (1/J_h)^{\frac{1}{m-1}}}, \end{aligned} \quad (3.22)$$

substituindo a equação (3.22) em (3.21), obtemos que um extremo de J_k é alcançado quando os graus de pertinência *fuzzy* u_{ki} , são da forma

$$\begin{aligned} u_{ki} &= \frac{1}{\sum_{h=1}^c (1/J_h)^{\frac{1}{m-1}}} \cdot \frac{1}{(J_k)^{\frac{1}{m-1}}} = \frac{1}{\sum_{h=1}^c \left(\frac{J_k}{J_h}\right)^{\frac{1}{m-1}}} \\ u_{ki} &= \left[\sum_{h=1}^c \left(\frac{J_k}{J_h}\right)^{\frac{1}{m-1}} \right]^{-1} = \left[\sum_{h=1}^c \left(\frac{\|\Phi(\mathbf{x}_i) - \mathbf{v}_k^\Phi\|^2}{\|\Phi(\mathbf{x}_i) - \mathbf{v}_h^\Phi\|^2} \right)^{\frac{1}{m-1}} \right]^{-1}. \end{aligned}$$

□

Temos ainda que,

$$\frac{\partial J_k}{\partial u_{ki}} = m(u_{ki})^{m-1} J_k \Rightarrow \frac{\partial^2 J_k}{\partial (u_{ki})^2} = m(m-1)(u_{ki})^{m-2} J_k,$$

com isso

$$\frac{\partial^2 J_k}{\partial (u_{ki}) \partial (u_{kj})} = 0 \quad \forall j \neq i.$$

Chegamos então, que a matriz Hessiana de J_k avaliada em $\mathbf{u}_k = (u_{k1}, \dots, u_{kn})$ pode ser escrita como

$$\mathbf{H}(\mathbf{u}_k) = \begin{pmatrix} \frac{m(m-1)J_1}{\sum_{h=1}^c \left(\frac{J_1}{J_h}\right)^{\frac{m-2}{m-1}}} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{m(m-1)J_c}{\sum_{h=1}^c \left(\frac{J_c}{J_h}\right)^{\frac{m-2}{m-1}}} \end{pmatrix},$$

além disso, obtemos que $\mathbf{H}(\mathbf{u}_k)$ é definida positiva, de modo que podemos concluir que esse extremo é um mínimo.

Os algoritmos dos métodos de agrupamento *fuzzy* no espaço de características baseados no *kernel* de Mahalanobis com matrizes de covariâncias globais e/ou locais completas e/ou diagonais foram executados de acordo com os seguintes passos:

1. Inicialização

Fixe K (o número de grupos), $2 \leq K < n$; fixe m , $1 < m < \infty$; fixe T (o número máximo de iterações); inicialize os graus de pertinência difusos u_{ki} ($k = 1, \dots, c, i = 1, \dots, n$) de tal forma que $u_{ki} \geq 0 \forall k, i$ e $\sum_{k=1}^c u_{ki} = 1, \forall i$; escolha uma função *kernel* adequada $\mathcal{K}$ e calcule a matriz *kernel* $\mathbf{K}$; faça $t = 1$.

2. **Etapa 1:** Definição das matrizes de covariâncias. Atualize o cálculo das matrizes $\mathbf{M}$ (globais) e $\mathbf{M}_k$ (locais) de acordo com as Eqs.((3.14),(3.16),(3.17), (3.18)).

3. **Etapa 2:** Definição da melhor partição difusa. Atualize os graus de pertinência difusos u_{ki} de acordo com a Eq.(3.20).

4. Critério de parada

Se $|J^{t+1} - J^t| \leq \varepsilon$ com $\varepsilon > 0$ ou $t > T$, PARE, caso contrário, faça $t = t + 1$ e volte ao passo 2.

Capítulo 4

Experimentos computacionais

Neste capítulo são apresentados os experimentos computacionais executados considerando os métodos de agrupamento *fuzzy* no espaço de características baseado no *Kernel* de Mahalanobis com distâncias quadráticas adaptativas introduzidos no Capítulo 3. Para tanto, os métodos propostos foram comparados a métodos clássicos de agrupamento como o *fuzzy k*-médiãs e suas versões baseadas no *kernel* Gaussiano e aos métodos propostos em [26], através de experimentos de Monte Carlo considerando quatro configurações de conjuntos de dados simulados. Além disso, foram realizadas aplicações em conjuntos de dados reais obtidos do Repositório de Aprendizagem de Máquina da Universidade da Califórnia, Irvine, Estados Unidos [2].

4.1 Índices de Avaliação

Para comparar os métodos de agrupamento considerados neste trabalho, utilizamos o Índice Corrigido de Rand (CR) e a taxa total de erro de classificação (OERC).

Seja $\mathcal{P} = \{\mathcal{P}_1, \dots, \mathcal{P}_i, \dots, \mathcal{P}_c\}$ a partição *a priori* de $\Omega = \{1, \dots, n\}$ em c classes e seja $P = \{P_1, \dots, P_k, \dots, P_K\}$ uma partição rígida de $\Omega = \{1, \dots, n\}$ em K grupos fornecidos por um algoritmo de agrupamento. Nesta perspectiva, as quantidades n_{ik} , $i = 1, \dots, c, k = 1, \dots, K$, representam o número de observações que estão na classe $\mathcal{P}_i$ e no grupo P_k e podem ser representadas na forma da Tabela 4.1, denominada matriz de confusão.

Tabela 4.1: Matriz de confusão

Classes	Grupos					
	P_1	$\dots$	P_k	$\dots$	P_K	$\sum$
$\mathcal{P}_1$	n_{11}	$\dots$	n_{1k}	$\dots$	n_{1K}	$n_{1\bullet} = \sum_{k=1}^K n_{1k}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
$\mathcal{P}_i$	n_{i1}	$\dots$	n_{ik}	$\dots$	n_{iK}	$n_{i\bullet} = \sum_{k=1}^K n_{ik}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
$\mathcal{P}_c$	n_{c1}	$\dots$	n_{ck}	$\dots$	n_{cK}	$n_{c\bullet} = \sum_{k=1}^K n_{ck}$
$\sum$	$n_{\bullet 1} = \sum_{i=1}^c n_{i1}$	$\dots$	$n_{\bullet k} = \sum_{i=1}^c n_{ik}$	$\dots$	$n_{\bullet K} = \sum_{i=1}^c n_{iK}$	$n = \sum_{i=1}^c \sum_{k=1}^K n_{ik}$

O Índice Corrigido de Rand (CR) é obtido como

$$CR = \frac{\sum_{i=1}^c \sum_{k=1}^K \binom{n_{ik}}{2} - \binom{n}{2}^{-1} \sum_{i=1}^c \binom{n_{i\bullet}}{2} \sum_{k=1}^K \binom{n_{\bullet k}}{2}}{\frac{1}{2} \left[\sum_{i=1}^c \binom{n_{i\bullet}}{2} + \sum_{k=1}^K \binom{n_{\bullet k}}{2} \right] - \binom{n}{2}^{-1} \sum_{i=1}^c \binom{n_{i\bullet}}{2} \sum_{k=1}^K \binom{n_{\bullet k}}{2}}, \quad (4.1)$$

onde $\binom{n}{2} = \frac{n(n-1)}{2}$, n_{ik} representa o número de observações que estão na classe $\mathcal{P}_i$ e no grupo P_k , $n_{i\bullet}$ representa o número de observações na classe $\mathcal{P}_i$, $n_{\bullet k}$ representa o número de observações no grupo P_k e n é o número total de observações no conjunto de dados [20] .

O CR avalia o grau de concordância (similaridade) entre uma partição *a priori* e uma partição fornecida por um método de agrupamento. Além disso, o CR não é sensível ao número de classes nas partições ou à distribuição das observações nos grupos. Finalmente, o CR assume valores no intervalo $[-1, 1]$, no qual o valor 1 indica concordância perfeita entre as partições, enquanto que valores próximos de zero ou negativos correspondem a concordância entre partições encontrada ao acaso [24].

Em problemas de classificação, cada grupo P_k é associado a uma classe *a priori* $\mathcal{P}_i$ e esta associação deve ser interpretada como se a verdadeira classe *a priori* fosse $\mathcal{P}_i$. Dessa forma, para uma observação pertencente a um dado grupo P_k a decisão está correta se a classe *a priori* dessa observação é $\mathcal{P}_i$. Para obter uma taxa de erro de classificação mínima, precisamos encontrar uma regra de decisão que minimize a probabilidade de erro.

Seja $\ell(\mathcal{P}_i, P_k)$ a probabilidade *a posteriori* de que uma observação pertença à classe $\mathcal{P}_i$ quando associada ao grupo P_k . Seja $\ell(P_k)$ a probabilidade de que a observação pertença ao grupo P_k . A função ℓ é conhecida como função de verossimilhança. A estimativa da máxima probabilidade *a posteriori* é a moda da probabilidade *a*

posteriori $\ell(\mathcal{P}_i, P_k)$ e o índice da classe *a priori* associado a esta moda é dada por:

$$MAP(P_k) = \arg \max_{1 \leq i \leq c} \ell(\mathcal{P}_i, P_k).$$

A regra de decisão de Bayes que minimiza a probabilidade média de erro é selecionar a classe *a priori* que maximiza a probabilidade *a posteriori*. A taxa de erro de classificação do grupo P_k é igual a $1 - \ell(\mathcal{P}_{MAP(P_k)}/P_k)$ e a taxa total de erro de classificação OERC é igual a

$$OERC = \sum_{k=1}^K \ell(P_k)(1 - \ell(\mathcal{P}_{MAP(P_k)}/P_k)).$$

Para uma amostra,

$$\ell(\mathcal{P}_{MAP(P_k)}/P_k) = \max_{1 \leq i \leq c} n_{ik}/n_{\bullet k}.$$

A taxa total de erro de classificação (OERC) foi concebida de modo a medir a habilidade de um algoritmo de agrupamento encontrar as classes *a priori* presentes em um conjunto de dados e é calculada da forma [6]:

$$OERC = \sum_{k=1}^K \frac{n_{\bullet k}}{n} \left(1 - \max_{1 \leq i \leq c} n_{ik}/n_{\bullet k} \right) = 1 - \frac{\sum_{k=1}^K \max_{1 \leq i \leq c} n_{ik}}{n}. \quad (4.2)$$

O índice OERC assume valores no intervalo $[0, 1]$, no qual valores próximos de zero indicam maior habilidade de um algoritmo na detecção de classe *a priori*.

4.2 Conjunto de Dados Simulados

Para cada conjunto de dados sintéticos simulamos utilizamos classes com tamanhos e formas diferentes. Os conjuntos de dados simulados tiveram 500 pontos cada, divididos em quatro classes com tamanhos ($n = 50, 100, 150$ e 200). Cada classe nesses dados foi desenhada de acordo com uma distribuição normal bivariada com vetor de médias e matriz de covariâncias representada por

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \text{ e } \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{bmatrix}.$$

Para tanto, consideramos quatro configurações de conjuntos de dados simulados:

1. As matrizes de covariâncias dos grupos são diagonais e aproximadamente iguais;

2. As matrizes de covariâncias dos grupos são diagonais e diferentes para cada grupo;
3. As matrizes de covariâncias dos grupos são completas e aproximadamente iguais;
4. As matrizes de covariâncias dos grupos são completas e diferentes para cada grupo.

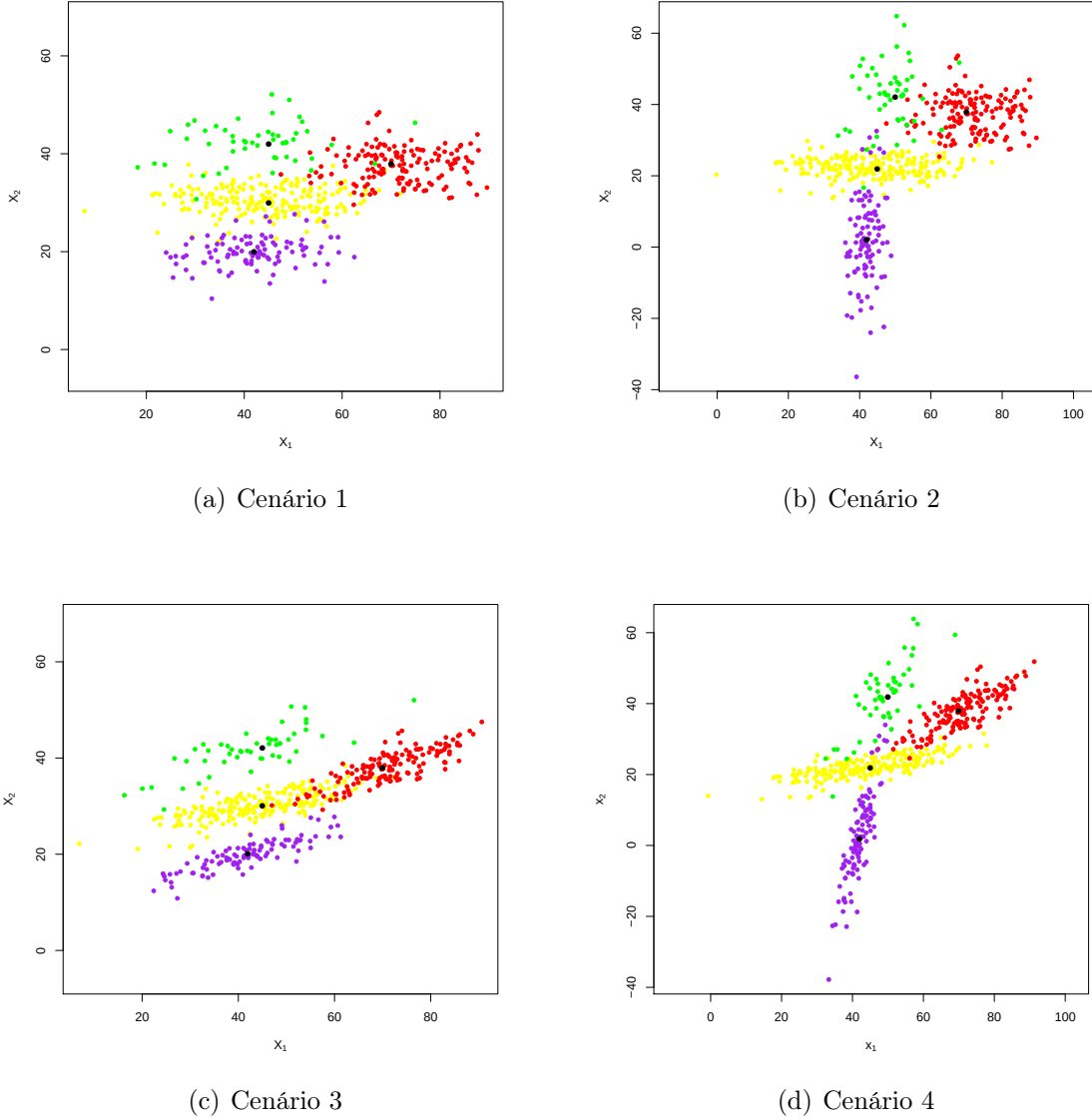
Na Tabela 4.2, apresentamos as configurações dos dados simulados para cada classe e seus respectivos parâmetros, bem como os valores dos vetores de média e as matrizes de covariâncias (composta pelas variâncias e as correlações de cada classe) para cada conjunto de dados simulados.

Tabela 4.2: Configurações dos cenários						
Cenários	Classes	μ_1	μ_2	σ_1^2	σ_2^2	ρ
1	1	45	30	100	9	0.0
	2	70	38	81	16	0.0
	3	45	42	100	16	0.0
	4	42	20	81	9	0.0
2	1	45	22	144	9	0.0
	2	70	38	81	36	0.0
	3	50	42	36	81	0.0
	4	42	20	9	144	0.0
3	1	45	30	100	9	0.7
	2	70	38	81	16	0.8
	3	45	42	100	16	0.7
	4	42	20	81	9	0.8
4	1	45	22	144	9	0.7
	2	70	38	81	36	0.8
	3	50	42	36	81	0.7
	4	42	20	9	144	0.8

Podemos observar na figura 4.1, as configurações dos quatros conjuntos de dados simulados considerados:

- (a) Matrizes de covariâncias diagonais e globais (quase iguais);
- (b) Matrizes de covariâncias diagonais e locais(diferente para cada grupo);
- (c) Matrizes de covariâncias completas e globais;
- (d) Matrizes de covariâncias completas e locais.

Figura 4.1: Conjunto de Dados Simulados



4.3 Avaliação com Dados Simulados

Nesta etapa, o desempenho dos métodos foram avaliados tomando por referência o Índice Corrigido de Rand (CR) e a taxa total de erro de classificação (OERC). Para isso, os índices CR e OERC foram calculados baseados na partição *fuzzy* obtida de uma simulação de Monte Carlo com 100 repetições para cada configuração de dados. Isso significa que cada replicação retornou uma matriz de confusão, sendo que os grupos obtidos pelo algoritmo de agrupamento foram colocados em linhas e as classes *a priori* foram colocadas em colunas.

Em cada replicação, os algoritmos de agrupamento foram executados 100 vezes até convergir para um valor estacionário do critério de adequação, e depois disso o melhor resultado para cada algoritmo foi selecionado de acordo com o critério

de adequação. Para tanto, os algoritmos foram aplicados aos conjuntos de dados simulados 1, 2, 3 e 4, obtendo 4 grupos. As partições dos 4 grupos fornecidas pelos algoritmos de agrupamento foram comparadas a partição *a priori* conhecida de 4 classes. Além disso, a média e o desvio padrão destas medidas foram computados.

No processo de análise dos dados simulados, procedemos da seguinte forma, para cada conjunto de dados simulados identificamos quais métodos mostraram melhor desempenho, tanto os métodos desenvolvidos por [26] como os propostos neste trabalho.

A tabela 4.3 apresenta o desempenho dos algoritmos KFCM.KM.GD, KFCM.KM.GF, KFCM.KM.LD e KFCM.KM.LF desenvolvidos por [26] e KFCM.FS.GD, KFCM.FS.GF, KFCM.FS.LD, KFCM.FS.LF propostos neste trabalho em valores de média e desvio padrão para o conjunto de dados simulados 1, de acordo com os índices CR e OERC. Ainda na tabela 4.3, e nas seguintes, é apresentado o desempenho do método *fuzzy c*-médias (FCM), assim como sua versão *kernel* baseada em kernelização da métrica (KFCM.KM) e no espaço de características (KFCM.FS).

Tabela 4.3: Desempenho dos algoritmos de agrupamento *fuzzy* no conjunto de dados simulados 1: média e desvio-padrão (entre parênteses) dos índices CR e OERC.

Autor(es)	Algoritmos	CR	OERC
SILVA (2018)	FCM	0.303 (0.050)	0.418 (0.044)
	KFCM.KM	0.321 (0.060)	0.405 (0.049)
	KFCM.KM.GD	0.623 (0.105)	0.197 (0.081)
	KFCM.KM.GF	0.617 (0.095)	0.198 (0.072)
	KFCM.KM.LD	0.588 (0.096)	0.227 (0.078)
Deste trabalho	KFCM.KM.LF	0.554 (0.096)	0.246 (0.075)
	KFCM.FS	0.341 (0.058)	0.379 (0.047)
	KFCM.FS.GD	0.716 (0.050)	0.122 (0.035)
	KFCM.FS.GF	0.649 (0.111)	0.172 (0.074)
	KFCM.FS.LD	0.687 (0.091)	0.140 (0.067)
	KFCM.FS.LF	0.632 (0.096)	0.179 (0.071)

De acordo com a tabela 4.3, temos que os algoritmos (KFCM.FS.GD) com distância quadrática definida por uma matriz simétrica diagonal definida positiva $\mathbf{M}$, que é a mesma para todos os grupos e (KFCM.FS.LD) com distância quadrática definida por uma matriz simétrica diagonal definida positiva $\mathbf{M}_k$, que é diferente para cada grupo, superaram os demais métodos.

Analisando esses dois algoritmos que apresentaram melhor desempenho, o algoritmo com matriz de covariâncias diagonal global (KFCM.FS.GD) demonstrou melhor explanação do conjunto de dados simulados 1, com o maior CR (0.716) e o menor OERC (0.122).

Dentre os algoritmos desenvolvidos por [26], destaca-se que (KFCM.KM.GD) apresentou melhor desempenho, com maior CR (0.623) e o menor OERC (0.197). Contudo, esse resultado ainda se mostra inferior aos obtidos pelos algoritmos (KFCM.FS.GD) e (KFCM.FS.LD).

Para o conjunto de dados simulados 2, o desempenho dos métodos de agrupamentos *fuzzy* é apresentado na tabela 4.4. Verifica-se que os algoritmos baseados em distâncias adaptativas locais, ou seja, que possuem matriz diagonal e completa de covariâncias distintas para cada grupo, (KFCM.FS.LD) e (KFCM.FS.LF), forneceram melhores resultados do que os demais algoritmos. Destaque para o (KFCM.FS.LD) com maior CR (0.768) e o menor OERC (0.106).

Tabela 4.4: Desempenho dos algoritmos de agrupamento *fuzzy* no conjunto de dados simulados 2: média e desvio-padrão (entre parênteses) dos índices CR e OERC.

Autor(es)	Algoritmos	CR	OERC
SILVA (2018)	FCM	0.397 (0.034)	0.345 (0.030)
	KFCM.KM	0.398 (0.038)	0.339 (0.037)
	KFCM.KM.GD	0.415 (0.033)	0.329 (0.029)
	KFCM.KM.GF	0.401 (0.032)	0.314 (0.036)
	KFCM.KM.LD	0.535 (0.062)	0.244 (0.049)
Deste trabalho	KFCM.KM.LF	0.467 (0.097)	0.293 (0.072)
	KFCM.FS	0.553 (0.079)	0.226 (0.063)
	KFCM.FS.GD	0.577 (0.098)	0.209 (0.066)
	KFCM.FS.GF	0.532 (0.078)	0.264 (0.064)
	KFCM.FS.LD	0.768 (0.039)	0.106 (0.029)
	KFCM.FS.LF	0.743 (0.050)	0.130 (0.045)

De acordo com o explanado na tabela 4.4, o algoritmo (KFCM.FS.LF) alcançou resultado inferior ao (KFCM.FS.LD), no entanto, mostrou eficiência superior ao método (KFCM.KM.LD), que apresentou melhor desempenho dentre os algoritmos desenvolvidos por [26] no conjunto de dados simulados 2. Corroborando com o visto no conjunto de dados simulados 1, os métodos propostos neste trabalho apresentaram melhor desempenho, em relação a contrapartida.

O desempenho dos métodos de agrupamento *fuzzy* para o conjunto de dados simulados 3 é apresentado na tabela 4.5. Neste cenário, dentre os algoritmos que explicaram melhor o conjuntos de dados simulados 3, o algoritmo (KFCM.FS.LF) se destacou, possuindo o maior CR (0.760) e o menor OERC (0.075). Contudo, o algoritmo (KFCM.KM.GF) desenvolvido por [26] obteve resultados próximos, com maior CR (0.758) e o menor OERC (0.083).

Infere-se então, que para este cenário, os métodos que apresentaram matriz de covariâncias completa, foram superior aos demais métodos. Este fato é verificado com o resultado dos métodos (KFCM.FS.LF), com maior CR (0.760) e o menor

Tabela 4.5: Desempenho dos algoritmos de agrupamento *fuzzy* no conjunto de dados simulados 3: média e desvio-padrão (entre parênteses) dos índices CR e OERC.

Autor(es)	Algoritmos	CR	OERC
SILVA (2018)	FCM	0.252 (0.028)	0.446 (0.033)
	KFCM.KM	0.264 (0.042)	0.441 (0.041)
	KFCM.KM.GD	0.452 (0.029)	0.311 (0.022)
	KFCM.KM.GF	0.758 (0.031)	0.083 (0.012)
	KFCM.KM.LD	0.435 (0.052)	0.328 (0.036)
Deste trabalho	KFCM.KM.LF	0.756 (0.029)	0.084 (0.011)
	KFCM.FS	0.251 (0.031)	0.434 (0.033)
	KFCM.FS.GD	0.540 (0.110)	0.227 (0.097)
	KFCM.FS.GF	0.755 (0.037)	0.080 (0.015)
	KFCM.FS.LD	0.483 (0.153)	0.268 (0.130)
	KFCM.FS.LF	0.760 (0.032)	0.075 (0.012)

OERC (0.075) e (KFCM.KM.GF), com maior CR (0.758) e o menor OERC (0.083).

Tabela 4.6: Desempenho dos algoritmos de agrupamento *fuzzy* no conjunto de dados simulados 4: média e desvio-padrão (entre parênteses) dos índices CR e OERC.

Autor(es)	Algoritmos	CR	OERC
SILVA.(2018)	FCM	0.416 (0.028)	0.329 (0.025)
	KFCM.KM	0.430 (0.025)	0.322 (0.022)
	KFCM.KM.GD	0.433 (0.024)	0.325 (0.022)
	KFCM.KM.GF	0.422 (0.026)	0.328 (0.019)
	KFCM.KM.LD	0.527 (0.037)	0.267 (0.028)
Deste trabalho	KFCM.KM.LF	0.783 (0.124)	0.099 (0.078)
	KFCM.FS	0.419 (0.026)	0.316 (0.023)
	KFCM.FS.GD	0.416 (0.027)	0.322 (0.024)
	KFCM.FS.GF	0.400 (0.071)	0.306 (0.056)
	KFCM.FS.LD	0.647 (0.108)	0.176 (0.073)
	KFCM.FS.LF	0.840 (0.025)	0.054 (0.010)

A tabela 4.6 apresenta o desempenho dos métodos de agrupamento *fuzzy* para o conjunto de dados simulados 4. Depreende-se neste cenário, que os métodos formulados com matriz de covariâncias completa (não diagonal) local, foram superior aos demais métodos. Nesta situação, o algoritmo (KFCM.FS.LF) mostrou melhor desempenho, com o maior CR (0.840) e o menor OERC (0.054). Dentre os métodos desenvolvidos por [26], o algoritmo que mais se aproximou do (KFCM.FS.LF) foi o (KFCM.KM.LF), com maior CR (0.783) e o menor OERC (0.099).

De acordo com a tabela 4.7, no cenário 1, os algoritmos baseados em matriz de covariâncias diagonal global (matriz única), mostraram melhor desempenho na explicação dos dados. Para o cenário 2, os algoritmos baseados em matriz de covariâncias

Tabela 4.7: Algoritmos de agrupamento *fuzzy* com melhor desempenho no conjunto de dados simulados 1,2,3 e 4.

Autor(es)	Cenário 1	Cenário 2	Cenário 3	Cenário 4
SILVA (2018)	KFCM.KM.GD	KFCM.KM.LD	KFCM.KM.GF	KFCM.KM.LF
Deste trabalho	KFCM.FS.GD	KFCM.FS.LD	KFCM.FS.LF	KFCM.FS.LF

diagonal local (matriz diferente para cada grupo), mostraram melhor desempenho. No cenário 3, o método (KFCM.KM.GF) baseado em matriz de covariâncias completa global (matriz única) e (KFCM.FS.LF) baseado em matriz de covariâncias completa local (matriz diferente para cada grupo), alcançaram melhores resultados. Para o cenário 4, os algoritmos baseados em matriz de covariâncias completa local (matriz diferente para cada grupo), apresentam melhor explanação dos dados.

Verificamos com esta avaliação, que os algoritmos propostos neste trabalho, métodos de agrupamento *fuzzy* no espaço de características baseado no *Kernel* de Mahalanobis com distâncias quadráticas adaptativas, para os conjuntos de dados simulados 1, 2 3 e 4, apresentaram desempenho superior aos demais métodos desenvolvidos por [26].

4.4 Conjunto de Dados Reais

Neste ambiente, foram realizados experimentos com nove conjuntos de dados reais obtidos do Repositório de Aprendizagem de Máquina da Universidade da Califórnia, denominados *Breast Tissue*, *Vertebral Column*, *Ecoli*, *Íris plant*, *Rukiln*, *Sonar*, *Thyroid Disease*, *Wisconsin diagnostic breast cancer (WDBC)* e *Wine*.

Na Tabela 4.8 (na qual n representa o número de observações, p representa o número de variáveis e c representa o número de classes *a priori*) temos as dimensões dos conjuntos de dados reais considerados. Nesta seção, corroborando com [2] descreveremos algumas informações a cerca de cada conjunto de dado real.

Breast Tissue

Este conjunto de dados consiste em seis tipos (classes) do Tecido Mamário: fibroadenoma, mastopatia, glandular, conectivo, adiposo e carcinoma. As seis classes foram medidas em frequência de KHz e plotadas no plano real, constituindo o espectro de impedância a partir do qual as características do tecido mamário são computadas.

Tabela 4.8: Dimensões dos conjuntos de dados reais

Conjunto de Dados	n	p	c
<i>Breast Tissue</i>	106	9	6
<i>Vertebral Column</i>	310	6	3
<i>Ecoli</i>	336	5	8
<i>Íris plant</i>	150	4	3
<i>Rukiln</i>	27	10	2
<i>Sonar</i>	208	41	2
<i>Thyroid Gland</i>	215	5	3
<i>WDBC</i>	569	30	2
<i>Wine</i>	178	13	3

Vertebral Column

Os dados neste conjunto foram organizados em duas tarefas de classificação diferentes, mas relacionadas. A primeira tarefa consiste em classificar os pacientes como pertencentes a uma das três classes: Normal (100 pacientes), Hernia do disco (60 pacientes) ou Espondilolistese (150 pacientes). Para a segunda tarefa, as classes Disk Hernia e Spondylolisthesis foram incorporadas em uma única classe rotulada como "anormal". Assim, a segunda tarefa consiste em classificar os pacientes como pertencentes a uma das duas classes: Normal (100 pacientes) ou anormal (210 pacientes). Cada paciente é representado no conjunto de dados por seis atributos biomecânicos derivados da forma e orientação da pélvis e da coluna lombar (nesta ordem): incidência pélvica, inclinação pélvica, ângulo de lordose lombar, inclinação sacra, raio pélvico e grau de espondilolistese.

Ecoli

Este conjunto de dados consiste em oito tipos (classes) de proteínas: lipoproteínas de membrana interna, lipoproteínas de membrana externa, proteínas da membrana interna com sequência de sinal clivável, outras proteínas da membrana externa, proteínas periplasmáticas, brãia interna proteínas com uma sequência de sinal não-clivável, proteínas da membrana interna sem uma sequência sinal e proteínas citoplasmáticas. Foram considerados quatro locais de localização da proteína na célula, na regiões do citoplasma, da membrana interna (citoplasmática), do periplasma e a da membrana externa.

Íris Plant

Este conjunto de dados consiste em três tipos (classes) de plantas da íris: íris setosa, íris versicolor e íris virgínica. As três classes têm 50 instâncias cada. Uma

classe é linearmente separável das outras duas; o último não é linearmente separável um do outro. Cada classe é descrita por quatro atributos de valor real. Esses atributos são: comprimento sepal, largura sepal, comprimento da pétala e largura da pétala.

Rukiln

Este conjunto de dados consiste em dois tipos (classes) de esmaltes de forno RU: corpo de porcelana oficial de Forno-RU e corpo de porcelana de Forno-Ru, com 19 e 8 casos, respectivamente. Cada instância é descrita por 10 variáveis de valor real que representam os elementos dos oligoelementos dos dois tipos de esmaltes do Forno-Ru. Esses elementos (variáveis) são: (1) Cu, (2) Sr, (3) Zn/Cu, (4) Ni, (5) Zn, (6) Ga, (7) Rb, (8) Zr, (9) Pb, and (10) Rb/Ga.

Sonar

Este conjunto de dados consiste em dois tipos (classes): sonar das minas e o sonar das rochas. O sonar da minas contém 111 observações obtidos por sinais de sonar captados em um cilindro de metal em vários ângulos e sob várias condições. O sonar das rochas contém 97 observações obtidos a partir de rochas em condições semelhantes. O sinal do sonar é transmitido para um chip modulador de frequência. O conjunto de dados contém sinais obtidos a partir de uma variedade de ângulos de aspecto diferentes, abrangendo 90 graus para o cilindro e 180 graus para a rocha.

Thyroid Gland

Este conjunto de dados consiste em três classes relativas ao estado da tireóide: normal, hipertireoidismo e hipotireoidismo. As classes (1, 2 e 3) têm, respectivamente, 150, 35 e 30 instâncias. Cada classe é descrita por cinco atributos de valor real. Esses atributos são: teste de absorção de resina T3, tiroxina sérica total, triiodotironina sérica total, hormônio basal de estimulação tireóidea (TSH) e diferença absoluta absoluta de valor de TSH.

Wisconsin diagnostic breast cancer (WDBC)

Neste conjunto de dados as características são calculadas a partir de uma imagem digitalizada de um aspirador com uma agulha fina (FNA) de uma massa do peito. Eles descrevem características dos núcleos celulares presentes na imagem. O plano de separação descrito acima foi obtido usando o método-árvore Multisurface (MSM-T), um método de classificação que usa programação linear para construir uma árvore

de decisão. As características relevantes foram selecionadas usando uma pesquisa exaustiva no espaço de 1-4 características e 1-3 separando planos.

Wine

Este conjunto de dados consiste em três tipos (classes) de vinhos cultivados na mesma região na Itália, mas derivados de três plantações diferentes. As classes (1, 2 e 3) têm, respectivamente, 59, 71 e 48 instâncias. Cada vinho é descrito por 13 atributos de valor real que representam as quantidades de 13 constituintes encontrados em cada um dos três tipos de vinhos. Esses atributos são: (1) álcool; (2) ácido málico; (3) cinzas; (4) alcalinidade das cinzas; (5) magnésio; (6) fenóis totais; (7) flavonóides; (8) fenóis não flavonóides; (9) proantocianinas; (10) intensidade de cor; (11) matiz; (12) OD280 / OD315 de vinhos diluídos e (13) prolina.

4.5 Avaliação com Dados Reais

Nesta avaliação, o desempenho dos métodos desenvolvidos por [26] e os propostos neste trabalho foram aplicados aos conjuntos de dados reais destacados na tabela 4.8 e analisados tomando por referência o Índice Corrigido de Rand (CR) e a taxa total de erro de classificação (OERC). Em cada caso, os algoritmos de agrupamento *fuzzy* foram executados 100 vezes (até a convergência para um valor estacionário do critério de adequação entre grupos), e depois disso o melhor resultado para cada algoritmo foi selecionado.

O desempenho dos métodos de agrupamento *fuzzy* para o conjunto de dados reais, de acordo com os índices CR e OERC, é apresentado na Tabela 4.9. Para cada conjunto de dados reais foram destacados os algoritmos que obtiveram melhor desempenho perante os demais (estão descritos em negrito), isto é, os algoritmos que apresentaram maior índice CR e menor OERC.

De acordo com a tabela 4.9, infere-se que os métodos propostos nesta dissertação apresentaram desempenho, em sua maioria, superior ou equivalente aos desenvolvidos por [26]. Dessa forma, realizando uma análise dos resultados para cada conjunto de dados reais, de acordo com os métodos, obteve-se que o (KFCM.FS.GD) mostrou desempenho superior aos demais métodos nos seguintes conjuntos de dados: *Breast Tissue*, *Íris plant*, contudo, no conjunto de dados *Wine* alcançou desempenho equivalente ao método (KFCM.KM.GD). De forma análoga, o método (KFCM.FS.LD) mostrou maior desempenho nos conjuntos de dados: *Rukiln* e *Thyroid Gland*, já para o conjunto de dados *Íris plant*, obteve resultado equivalente aos métodos (KFCM.FS.GD) e (KFCM.FS.GF), no conjunto *WDBC*, igualhou-se ao método (KFCM.KM.LD). Para o conjunto de dados *Vertebral Column*, os algo-

ritmos (KFCM.KM) e (KFCM.FS) apresentaram melhor desempenho, com valores de CR e OERC próximos. Ainda no conjunto *Vertebral Column*, dentre os demais métodos propostos, destaca-se o comportamento do algoritmo (KFCM.FS.GF), que mostrou um desempenho próximo ao obtido pelo (KFCM.FS).

Para o conjunto de dados *Ecoli*, o método (KFCM.FS.LF) demonstrou melhor desempenho em relação aos demais métodos propostos neste trabalho, como também os desenvolvidos por [26]. Neste panorama, o método destacou-se ao apresentar maior CR (0.395) e o menor OERC (0.222). Para o conjunto de dados *Sonar*, verifica-se que o método (KFCM.FS) alcançou uma eficiência equivalente ao método clássico (FCM), com o maior CR (0.006) e o menor OERC (0.447). Em sequência, no conjunto de dados *Íris plant*, pode-se observar uma eficiência equivalente entre os métodos (KFCM.FS.GD), (KFCM.FS.GF) e (KFCM.FS.LD), com maior CR (0.868) e o menor OERC (0.044).

A partir dos valores descritos para o conjunto de dados reais e simulados, observa-se de forma geral a superioridade dos algoritmos propostos neste trabalho, contudo, para o conjunto de dados reais, obteve-se valores equivalentes para o desempenho dos métodos, este fato pode se verificado nos conjuntos de dados *Wine* e *WDBC*.

Na tabela 4.10, é apresentado um aparato geral da eficiência dos métodos desenvolvidos nesta dissertação, em cada conjunto de dados. Para esta avaliação, pretendeu-se analisar quais métodos dentre os desenvolvidos neste trabalho apresentaram melhor desempenho nos conjuntos de dados simulados e reais simultaneamente. Obteve-se então, que o algoritmo (KFCM.FS.GF) foi o menos eficiente, apresentando uma explicação satisfatória dos dados apenas para o conjunto de dados reais *Íris plant*. O algoritmo (KFCM.FS.LF) alcançou uma melhor explicação nos cenários 3 e 4 no conjunto de dados simulados e *Ecoli* para os dados reais. Já o algoritmo (KFCM.FS.GD), mostrou mais eficiência na explicação dos conjuntos *Breast Tissue*, *Íris plant* e *Wine*, para os dados reais, e o cenário 1, para os dados simulados. Por fim, o algoritmo (KFCM.FS.LD) explicou um maior número de conjuntos, sendo estes os conjuntos *Íris plant*, *Rukiln*, *Thyroid Gland* e *WDBC* dentre os dados reais e o cenário 2 para os dados simulados.

Avaliamos neste capítulo o desempenho dos métodos de agrupamento *fuzzy* no espaço de características baseado no *Kernel* de Mahalanobis com distâncias quadráticas adaptativas introduzidos no Capítulo 3. Os resultados numéricos dos experimentos de Monte Carlos evidenciam a superioridade dos algoritmos de agrupamentos *fuzzy* propostos neste trabalho. Contudo, os experimentos com dados reais apresentados neste capítulo expõe o potencial da utilidade dos algoritmos propostos em diversas situações práticas da realidade.

Tabela 4.9: Desempenho dos métodos de agrupamento *fuzzy* nos conjuntos de dados reais dos índices CR e OERC.

<i>Algoritmos</i>	<i>Breast Tissue</i>		<i>Vertebral Column</i>		<i>Ecoli</i>	
	CR	OERC	CR	OERC	CR	OERC
FCM	0.125	0.613	0.304	0.287	0.367	0.488
KFCM.KM	0.169	0.556	0.326	0.309	0.338	0.524
KFCM.KM.GD	0.298	0.471	0.248	0.374	0.348	0.518
KFCM.KM.GF	0.258	0.491	0.311	0.316	0.336	0.524
KFCM.KM.LD	0.278	0.481	0.261	0.364	0.297	0.538
KFCM.KM.LF	0.189	0.518	0.120	0.458	0.387	0.422
KFCM.FS	0.169	0.556	0.327	0.320	0.395	0.223
KFCM.FS.GD	0.299	0.469	0.246	0.372	0.348	0.518
KFCM.FS.GF	0.257	0.491	0.315	0.318	0.336	0.524
KFCM.FS.LD	0.278	0.479	0.260	0.366	0.297	0.538
KFCM.FS.LF	0.187	0.520	0.120	0.458	0.395	0.222
<i>Algoritmos</i>	<i>Iris plant</i>		<i>Rukiln</i>		<i>Sonar</i>	
	CR	OERC	CR	OERC	CR	OERC
FCM	0.729	0.106	0.578	0.111	0.006	0.447
KFCM.KM	0.757	0.093	0.002	0.407	0.001	0.461
KFCM.KM.GD	0.868	0.046	0.137	0.296	-0.004	0.476
KFCM.KM.GF	0.867	0.045	0.206	0.259	-0.003	0.475
KFCM.KM.LD	0.866	0.044	0.709	0.074	-0.004	0.466
KFCM.KM.LF	0.743	0.100	0.284	0.222	-0.005	0.466
KFCM.FS	0.757	0.093	0.376	0.222	0.006	0.447
KFCM.FS.GD	0.868	0.044	0.137	0.296	-0.004	0.476
KFCM.FS.GF	0.868	0.044	0.206	0.259	-0.004	0.476
KFCM.FS.LD	0.868	0.044	0.720	0.070	-0.004	0.476
KFCM.FS.LF	0.743	0.100	0.284	0.222	-0.005	0.486
<i>Algoritmos</i>	<i>Thyroid Gland</i>		<i>WDBC</i>		<i>Wine</i>	
	CR	OERC	CR	OERC	CR	OERC
FCM	0.441	0.209	0.491	0.146	0.354	0.314
KFCM.KM	0.128	0.409	0.550	0.126	0.375	0.292
KFCM.KM.GD	0.152	0.409	0.773	0.059	0.835	0.056
KFCM.KM.GF	0.002	0.609	0.074	0.362	0.341	0.354
KFCM.KM.LD	0.877	0.037	0.818	0.047	0.742	0.089
KFCM.KM.LF	0.601	0.135	0.677	0.088	0.169	0.387
KFCM.FS	0.224	0.214	0.550	0.126	0.375	0.292
KFCM.FS.GD	0.152	0.409	0.773	0.059	0.835	0.056
KFCM.FS.GF	0.002	0.609	0.074	0.362	0.341	0.354
KFCM.FS.LD	0.880	0.034	0.818	0.047	0.742	0.089
KFCM.FS.LF	0.601	0.135	0.677	0.088	0.169	0.387

Tabela 4.10: Algoritmos de agrupamento *fuzzy* com melhor desempenho no conjunto de dados simulados e reais.

<i>Dados</i>	KFCM.FS.GF	KFCM.FS.LF	KFCM.FS.GD	KFCM.FS.LD
<i>Simulados</i>		Cenário 3 Cenário 4	Cenário 1	Cenário 2
<i>Reais</i>	<i>Íris plant</i>	<i>Ecoli</i>	<i>Breast Tissue</i> <i>Íris plant</i> <i>Wine</i>	<i>Íris plant</i> <i>Rukiln</i> <i>Thyroid Gland</i> <i>WDBC</i>

Capítulo 5

Conclusões

Neste trabalho, a partir da mudança das matrizes de covariâncias e da definição do *Kernel* adaptativo de Mahalanobis, introduzimos os métodos de agrupamento *fuzzy* no espaço de características baseado no *Kernel* de Mahalanobis com distâncias quadráticas adaptativas, no qual comparamos aos métodos de agrupamentos baseados em kernelização da métrica, desenvolvidos por [26]. Para tanto, recorrendo a ideia de minimização observada na otimização, determinamos uma expressão para obtenção dos melhores centróides dos grupos e para os graus de pertinência *fuzzy* por meio do *Kernel* de Mahalanobis. Determinamos também, as expressões que definem as matrizes de covariâncias *fuzzy* para cada método por meio da derivada de uma função objetivo (funcional) em relação a cada matriz de covariâncias.

Visando a verificação e utilidade dos algoritmos, os mesmos foram aplicados por meio de simulações de Monte Carlos a dois ambientes específicos, sendo estes, conjuntos de dados simulados e reais. Para os conjuntos de dados simulados, foram utilizados quatro cenários, com 500 pontos cada, com tamanhos ($n = 50, 100, 150$ e 200). Para os dados reais, foram utilizados nove conjuntos de dados do Repositório de Aprendizagem de Máquina da Universidade da Califórnia, Irvine, Estados Unidos, denominados *Breast Tissue*, *Vertebral Column*, *Ecoli*, *Íris plant*, *Rukiln*, *Sonar*, *Thyroid Gland*, *Wisconsin diagnostic breast cancer (WDBC)* e *Wine*. Para tanto, adotou-se o Índice Corrigido de Rand (CR) e a taxa total de erro de classificação (OERC), como forma de avaliar os métodos.

Com a realização dos experimentos numéricos no conjunto de dados simulados, comprovou-se que o método de agrupamento *fuzzy* baseado em distância quadrática adaptativa local, com matriz de covariâncias completa (KFCM.FS.LF), alcançou valores bem melhores que o conquistado pelos, até então, mais bem sucedidos algoritmos apresentados por [26]. Destaca-se que o algoritmo (KFCM.FS.LF) mostrou-se superior aos demais nos cenários 3 e 4.

Para o conjunto de dados reais, em uma visão geral, as melhores classificações de desempenho foram obtidas pelos algoritmos propostos neste trabalho. Contudo,

para o conjunto de dados *WDBC*, os métodos (KFCM.KM.LD) e (KFCM.FS.LD) obtiveram desempenho equivalente, o mesmo aconteceu no conjunto de dados *Wine*, com os métodos (KFCM.KM.GD) e (KFCM.FS.GD).

Ainda sobre o conjunto de dados reais, destaca-se a eficiência dos algoritmos propostos neste trabalho, em especial do método (KFCM.FS.GD) que apresentou melhor explanação em três conjuntos de dados, *Breast Tissue*, *Íris plant* e *Wine*, e (KFCM.FS.LD) que alcançou melhores resultados nos conjuntos de dados *Rukiln*, *Thyroid Gland*, *Íris plant* e *WDBC*.

Pelos resultados numéricos obtidos, comprovou-se a eficiência dos métodos de agrupamento com distâncias quadráticas adaptativas definidas por matriz de covariâncias diagonal global (KFCM.FS.GD) e local (KFCM.FS.LD) frente a outras versões. Em virtude dos resultados obtidos nesta dissertação de mestrado, indicamos a utilização destes métodos.

Para concluir, evidenciamos que os objetivos propostos neste trabalho foram alcançados. Nesta perspectiva, comprovou-se que os métodos de agrupamento *fuzzy* no espaço de características, baseado no *Kernel* de Mahalanobis com distâncias quadráticas adaptativas, definidas por matriz de covariâncias completa (não diagonal) e diagonal, única para todo grupo, como também, diferente para cada grupo, respectivamente, alcançaram valores bem melhores do que os métodos de agrupamento desenvolvidos por [26], quando avaliados por meio do CR e da OERC.

5.1 Sugestões para Trabalhos Futuros

Como parte dos estudos futuros, destacam-se os seguintes tópicos:

- Estender o trabalho para dados simbólicos de natureza intervalar;
- Estudar o uso de matrizes de *Kernel* aproximadas, para melhorar o desempenho computacional;
- Explorar o desenvolvimento de outras funções *Kernel* baseados em distâncias quadráticas.

Referências Bibliográficas

- [1] AIZERMAN, M. A., BRAVERMAN, E. A., ROZONOER, L., 1964, “Theoretical foundations of the potential function method in pattern recognition learning”. In: *Automation and Remote Control*, n. 25, Automation and Remote Control, pp. 821–837.
- [2] BACHE, K., LICHMAN, M., 2013. “UCI Machine Learning Repository”. Disponível em: <<http://archive.ics.uci.edu/ml>>.
- [3] BEN-HUR, A., HORN, D., SIEGELMANN, H. T., et al., 2001, “Support Vector Clustering”, *J. Mach. Learn. Res.*, v. 2, pp. 125–137.
- [4] BEZDEK, J. C., 1981, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA, Kluwer Academic Publishers. ISBN: 0306406713.
- [5] BORER, S., GERSTNER, W., 2002, “A new kernel clustering algorithm”. In: *Proceedings of the Ninth Internat. Conf. on Neural Information Processing*, v. 5, pp. 2527–2531.
- [6] BREIMAN, L., FRIEDMAN, J., STONE, C. J., et al., 1984, *Classification and Regression Trees*. Boca Raton, Chapman & Hall/CRC.
- [7] CAMASTRA, F., VERRI, A., 2005, “A Novel Kernel Method for Clustering”, *IEEE Trans. Neural Networks*, v. 27, n. 5, pp. 801–804.
- [8] CAMPS-VALLS, G., RODRIGO-GONZÁLEZ, A., MUÑOZ-MARÍ, J., et al., 2007, “Hyperspectral image classification with mahalanobis relevance vector machines.” In: *IGARSS*, pp. 3802–3805.
- [9] CASSIANO, K. M., 2014, *Análise de Séries Temporais Usando Análise Espectral Singular (SSA) e Clusterização de Suas Componentes Baseada em Densidade*. Tese de doutorado em engenharia elétrica, Pontifícia Universidade Católica do Rio de Janeiro, PUC-Rio, Rio de Janeiro, RJ, Brasil.
- [10] DIDAY, E., SIMON, J. C., 1976, “Digital Pattern Classification”. cap. Clustering Analysis, pp. 47–94, Berlin, Springer.

- [11] FERREIRA, M. R. P., 2013, *Agrupamento baseado em kernel com ponderação automática das variáveis via distâncias adaptativas*. Tese (doutorado em ciência da computação), Universidade Federal de Pernambuco, UFPE, Recife, PE.
- [12] FERREIRA, M., DE CARVALHO, F., 2014, “A kernel k-means clustering algorithm based on an adaptive Mahalanobis kernel”. In: *Neural Networks (IJCNN), 2014 International Joint Conference on*, pp. 1885–1892, July. doi: 10.1109/IJCNN.2014.6889653.
- [13] FILIPPONE, M., CAMASTRA, F., MASULLI, F., et al., 2008, “A survey of kernel and spectral methods for clustering”, *Pattern Recognition*, v. 41, pp. 176–190.
- [14] GIROLAMI, M., 2002, “Mercer Kernel-based clustering in feature space”, *IEEE Trans. Neural Networks*, v. 13, n. 3, pp. 780–784.
- [15] GRAEPEL, T., OBERMAYER, K., 1998, “Fuzzy topographic kernel clustering”. In: Brauer, W. (Ed.), *Proceedings of the Fifth GI Workshop Fuzzy Neuro Systems '98*, pp. 90–97.
- [16] GRAVES, D., PEDRYCZ, W., 2010, “Kernel-based fuzzy clustering and fuzzy clustering: A comparative experimental study”, *Fuzzy Sets and Systems*, v. 161, pp. 522–543.
- [17] HAYKIN, S., 1998, *Neural Networks: A Comprehensive Foundation*. 2nd ed. Upper Saddle River, NJ, USA, Prentice Hall PTR. ISBN: 0132733501.
- [18] HOLBIG, C., DE SOUZA JUNIOR, P., MACHADO SAMPAIO, R., et al., 2013, “Uso da Linguagem R em Ambientes Paralelos”, (03).
- [19] HÖPPNER, F., KLAWONN, F., KRUSE, R., et al., 1999, *Fuzzy Cluster Analysis*. John Wiley & Sons, Inc. ISBN: 3-540-40317-5.
- [20] HUBERT, L., ARABIE, P., 1985, “Comparing partitions”, *Journal of Classification*, v. 2, pp. 193–218.
- [21] JAIN, A. K., 2010, “Data clustering: 50 years beyond K-means”, *Pattern Recogn. Lett.*, v. 31, n. 8 (jun.), pp. 651–666. ISSN: 0167-8655. doi: 10.1016/j.patrec.2009.09.011. Disponível em: <<http://dx.doi.org/10.1016/j.patrec.2009.09.011>>.
- [22] LINDEN, R., 2009, “Técnicas de agrupamento”, *Revista de Sistemas de Informação da FSMA*, v. 4, pp. 18–36.

- [23] MANLY, B. F. J., 2013, “Métodos Estatísticos Multivariados: uma introdução .” ISBN: 978-85-7780-185-5.
- [24] MILLIGAN, G. W., 1996, “Clustering and Classification”. cap. Clustering validation: results and implications for applied analysis, pp. 341–375, Singapore, World Scientific.
- [25] SCHÖLKOPF, B., SMOLA, A. J., MÜLLER, K. R., 1998, “Nonlinear component analysis as a kernel eigenvalue problem”, *Neural Comput.*, v. 10, n. 5, pp. 1299–1319.
- [26] SILVA, A. S., 2018, *Agrupamento Fuzzy Baseado no Kernel de Mahalanobis com Distâncias Quadráticas Adaptativas*. Dissertação (mestrado em modelagem matemática e computacional), Universidade Federal da Paraíba, UFPB, João Pessoa, PB, Brasil.
- [27] VALE, M. N., 2005, *Agrupamentos de dados: Avaliação de Métodos e Desenvolvimento de Aplicativo para Análise de Grupos*. Dissertação (mestre em ciências em engenharia elétrica), Pontifícia Universidade Católica do Rio de Janeiro, PUC-Rio, Rio de Janeiro, RJ, Brasil.