



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

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

Alisson dos Santos Silva

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 R. P. Ferreira

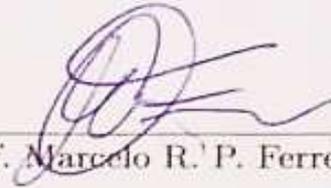
João Pessoa
Abril de 2018

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

Alisson dos Santos Silva

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 R. P. Ferreira, D.Sc.

Prof. Sérgio de Carvalho Bezerra, D.Sc.



Prof. Eufrásio de Andrade Lima Neto, D.Sc.

Prof. Hemílio Fernandes Campos Coêlho, D.Sc.

JOÃO PESSOA, PB – BRASIL
ABRIL DE 2018

Catálogo na publicação
Seção de Catalogação e Classificação

S586a Silva, Alisson Dos Santos.

AGRUPAMENTO FUZZY BASEADO NO KERNEL DE MAHALANOBIS COM
DISTÂNCIAS QUADRÁTICAS ADAPTATIVAS / Alisson Dos Santos
Silva. - João Pessoa, 2018.

61 f. : il.

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

1. Agrupamento Fuzzy. 2. Kernel de Mahalanobis. 3.
Distâncias Quadráticas Adaptativas. I. Ferreira,
Marcelo Rodrigo Portela. II. Título.

UFPB/BC

*O coração do homem planeja o
seu caminho, mas é o Senhor
que firma seus passos*
Provérbios 16, 9.

Agradecimentos

Agradeço primeiramente à nosso Senhor Jesus Cristo, por sempre ter guiado todos os meus passos, iluminando os meus caminhos e mostrando sempre a direção certa ao longo da minha vida. E a Nossa Senhora de Fátima, mãe e mestra que sempre intercede por minha vida.

Agradeço aos meus pais Antônio e Maria de Fátima, e a meu irmão Alan e meu sobrinho Arthur por todo amor, carinho, dedicação ao longo da minha história, incentivando e me apoiando em todos os momentos ao longo da minha formação acadêmica e profissional.

A minha eterna Avó, Maria Guedes ao qual deixo meu amor e carinho por sua eterna memória.

Aos meus familiares, as minhas tias, Alice, Margarida e Lindalva, ao meu tio e padrinho Luís e a minha madrinha Lourdes por todo amor, carinho, exemplo e dedicação durante toda a minha vida.

Aos professores Marcelo Rodrigo e Sérgio de Carvalho por terem me orientado nesses dois anos no mestrado. Em que gostaria de agradecer imensamente pela paciência e dedicação.

Agradeço a minha amiga Maizza Micaelle que considero como irmã, pela amizade, carinho e dedicação desde quando entrei na UFPB, me dando apoio em todos as minhas etapas profissionais e acadêmicas desde a minha graduação e agora também na pós-graduação.

Aos meus Amigos Luan, Geislane, Walissioneide e Raul por todo apoio, motivação e momentos de descontração durante este dois anos de pós-graduação.

Agradeço aos meus amigos e colegas da pós-graduação por estarem próximos nos momentos de estudo e de descontração, especialmente a Camila e a Josevandro. Aos professores do PPGMMC que foram tão importantes na minha formação acadêmica, por engrandecer meus conhecimentos ao longo desses dois anos. Ao CNPq, pelo apoio financeiro.

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* BASEADO NO *KERNEL* DE MAHALANOBIS COM DISTÂNCIAS QUADRÁTICAS ADAPTATIVAS

Alisson dos Santos Silva

Abril/2018

Orientador: Marcelo R. P. Ferreira

Programa: Modelagem Matemática e Computacional

Nesta dissertação de mestrado, são propostos métodos de agrupamentos fuzzy baseados no Kernel de Mahalanobis com distâncias quadráticas adaptativas definidas por matrizes de covariâncias diagonais comuns a todos os grupos, ou diferentes para cada grupo e matrizes de covariância completas comuns a todos os grupos ou diferentes para cada grupo. Este kernel foi construído a partir de uma distância quadrática adaptativa definida por uma matriz simétrica positiva-definida que é modificada a cada iteração do algoritmo que também será proposto. Os algoritmos propostos serão comparados com os diversos métodos de agrupamentos tradicionais conhecidos na literatura como o k-médias, o fuzzy k-médias e suas versões baseadas no Kernel Gaussiano. A utilidade dos métodos propostos foi demonstrada através de experimentos numéricos com conjuntos de dados simulados por meio de simulações de Monte Carlo e conjuntos de dados reais, cujo desempenho foi avaliado de acordo com os Índice Corrigido de Rand (CR) e a Taxa Total de Erro de Classificação (OERC). Esses métodos apresentaram, na maioria das situações experimentais consideradas nesta dissertação de mestrado, desempenhos superiores aos métodos de agrupamento clássico.

Palavras-chave: Agrupamento Fuzzy; Kernel de Mahalanobis; Distância Quadráticas Adaptativas.

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

THESIS TITLE

Alisson dos Santos Silva

April/2018

Advisor: Marcelo R. P. Ferreira

Program: Computational Mathematical Modelling

In this work, we propose fuzzy clustering methods based on a Mahalanobis-type kernel with adaptive quadratic distances defined either by fuzzy covariance matrices that are common to all clusters or by a fuzzy covariance matrix that can be different for each cluster. These matrices can be also full or diagonal. The Mahalanobis-type kernel functions were built from adaptive quadratic distances parameterized by symmetric, positive-defined matrices that changes at each iteration of the algorithm. The proposed algorithms were evaluated and compared with the main clustering methods in the literature as well as with their kernelized versions through experiments with simulated and real data.

Keywords: Fuzzy clustering, Mahalanobis kernel, Adaptive quadratic distances.

Sumário

Lista de Figuras	ix
Lista de Tabelas	x
1 Introdução	1
1.1 Objetivos	4
2 Revisão Bibliográfica	5
2.1 Técnicas Não-Hierárquicas	7
2.1.1 Método <i>c</i> -Médias	7
2.1.2 Método <i>Fuzzy c</i> -Médias	8
2.1.3 Fuzzy <i>c</i> -Médias baseado em Kernel	10
2.1.4 Kernel <i>c</i> -médias difuso baseado em Kernelização da métrica	12
2.1.5 Kernel <i>c</i> -médias difuso no espaço de características	13
3 Método Proposto	16
3.1 Introdução	16
3.2 Método de agrupamento fuzzy baseado no Kernel de Mahalanobis com distâncias quadráticas adaptativas	18
4 Esquema Experimental	28
4.1 Introdução	28
4.2 Índices de Avaliação	28
4.3 Conjunto de Dados Simulados	30
4.4 Avaliação com Dados Simulados	32
4.5 Conjuntos de Dados Reais	41
4.6 Avaliação com Dados Reais	44
5 Conclusões	46
Referências Bibliográficas	48

Lista de Figuras

4.1	Conjuntos de Dados Simulados.	32
4.2	Conjuntos de Dados Simulados do 1º Cenário.	34
4.3	Conjuntos de Dados Simulados do 2º Cenário.	36
4.4	Conjuntos de Dados Simulados do 3º Cenário.	38
4.5	Conjuntos de Dados Simulados do 4º Cenário.	40

Lista de Tabelas

2.1	Funções de Ligação	6
4.1	Matriz de confusão	29
4.2	Configurações dos Cenários	31
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.	33
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.	35
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.	37
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.	39
4.7	Dimensões dos conjuntos de dados reais.	41
4.8	Desempenho dos métodos de agrupamento fuzzy nos conjuntos de dados reais dos índices CR e OERC.	45

Capítulo 1

Introdução

Os métodos de agrupamentos englobam uma variedade de técnicas e algoritmos cujo objetivo é identificar padrões, alocando objetos, indivíduos, genes, etc que apresentam características semelhantes em grupos similares (JAIN [1]). Exemplos de aplicações incluem: segmentação de mercado, caracterização de regiões geográficas que possuem o mesmo desenvolvimento econômico, resultados de buscas na internet, agrupar páginas da internet com base em sua caracterização, classificar perfil de redes sociais com base em suas preferências ou características. A Análise de Agrupamento estuda todo um conjunto de relações interdependentes entre as variáveis. A finalidade dessa técnica é reunir os objetos (indivíduos, elementos) verificados nos grupos em que exista homogeneidade dentro do grupo e heterogeneidade entre os grupos, objetivando propor classificações (VICINI [2]). Popularmente essas técnicas podem ser divididas em métodos hierárquicos e métodos particionais (não-hierárquico).

Os métodos hierárquicos são técnicas multivariadas simples onde os dados são divididos sucessivamente até produzir uma estrutura de agrupamento hierarquizada. Um exemplo de saída de agrupamento que possuem uma estrutura hierárquica é o dendograma, também chamado de gráfico de árvore, ele é formado por agrupamento de observações em cada passo e em seus níveis de similaridade. O nível de similaridade é medido ao longo do eixo vertical e as diferentes observações são listadas ao longo do eixo horizontal. Os métodos hierárquicos possuem duas abordagens: os aglomerativos e os divisivos. Os aglomerativos começam com cada objeto ou observação em um grupo separado, em que os grupos mais parecidos são combinados para construir um novo grupo, o método é repetido até que todos os objetos sejam finalmente combinados em um único grupo (MANLY [3]). Os divisivos consideram que os objetos são todos colocados em um único grupo inicialmente, e então este é partido em dois grupos separando o objeto que está distante em média dos outros objetos. E em seguida um desses dois grupos é dividido, formando um total de três grupos. Esse procedimento continua até que todas as observações estejam em

grupos unitários, o oposto do método aglomerativo (MANLY [3]).

Os métodos particionais são mais flexíveis do que os hierárquicos e permitem que elementos amostrais mudem de grupo em cada passo do algoritmo se essa mudança permitir uma melhor solução em termos de variabilidade da partição resultante. Seu objetivo é buscar diretamente uma partição dos dados em k grupos de modo a preservar coesão interna e isolamento entre os grupos, cujo valor de k deve ser pré-definido (MINGOTI [4]). Os métodos de agrupamento particionais foram desenvolvidos sob dois diferentes paradigmas: agrupamento rígido *hard* (JAIN [1]) e agrupamento difuso *fuzzy* (HÖPPNER *et al.* [5]). Nos métodos de agrupamento do tipo rígido, os grupos são naturalmente disjuntos e não se sobrepõem. Nesse caso, cada padrão pode pertencer a um e somente um grupo. No caso dos métodos de agrupamento do tipo difuso, um padrão 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 (HÖPPNER *et al.* [5]).

Um componente importante de qualquer método de agrupamento é a medida de dissimilaridade (ou de similaridade). Alguns exemplos básicos são as medidas de distâncias, entre as mais comuns estão a Euclidiana, a de Mahalanobis e o Coeficiente de Correlação de Pearson. A distância Euclidiana é a mais comumente utilizada em métodos de agrupamento particionais (rígido e difuso). Segundo FERREIRA e DE CARVALHO [6] os métodos de agrupamento baseados na distância Euclidiana apresentam bom desempenho quando aplicados a conjuntos de dados nos quais os grupos são aproximadamente hipersféricos e aproximadamente linearmente separáveis. Contudo, quando a estrutura dos dados é complexa, i.e., grupos com formas não hipersféricas e/ou padrões não-linearmente separáveis, esses métodos podem não apresentar desempenho satisfatório. 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*.

Algoritmos de agrupamento rígido baseados em *kernel* foram propostos por diversos autores em que desenvolveram uma versão kernelizada do algoritmo SOM (Mapa Auto-Organizáveis). (ZHANG e CHEN [7], ZHANG *et al.* [8], CHEN e ZHANG [9], ZHANG e CHEN [10]. MACDONALD e FYFE [11], INOKUCHI e MIYAMOTO [12]). O Algoritmo SOM é usado como ferramenta de agrupamento para incorporar dados multidimensionais em espaços tipicamente bidimensionais para visualização de dados. Estes mapeamentos são capazes de preservar as relações de vizinhança dos dados de entrada, tornando assim aplicável em diversas áreas, como reconhecimento de voz, análise exploratória de dados e otimização combinatória. Além disso, versões kernelizadas dos algoritmos *mountain* e *neural gas* foram apresentadas, respectivamente, em (KIM *et al.* [13] e QINAND e SUGANTHAN [14]).

Diversos estudos demonstraram que os métodos de agrupamento baseados em

kernel possuem melhor desempenho quando aplicados na separação de grupos que apresentam características não lineares, ou seja, quando a estrutura dos conjuntos de dados é complexa e produzem superfícies que são parecidas topologicamente como elipse ou círculos (FILIPPONE *et al.* [15], CAMASTRA e VERRI [16], BEN-HUR *et al.* [17], GRAVES e PEDRYCZ [18], GIROLAMI [19], BEZDEK [20], KOHONEN [21]). Algoritmos baseados em *kernel* vêm sendo desenvolvidos sob duas abordagens principais: kernelização da métrica, onde os protótipos dos grupos são obtidos no espaço original dos padrões de entrada e as distâncias dos padrões aos protótipos dos grupos são calculadas através de funções *kernel*; e agrupamento no espaço de características, no qual os protótipos dos grupos são obtidos no espaço de características (FERREIRA e DE CARVALHO [6]).

A função *kernel* mede a similaridade entre dois exemplos em um espaço p -dimensional. Dentre as diferentes funções *kernel* existentes, o *kernel Gaussiano* é o mais comum. Esse *kernel* é baseado na distância Euclidiana e necessita apenas estimar um único parâmetro, além de fornecer bons resultados (FILIPPONE *et al.* [15], GRAVES e PEDRYCZ [18]). Os algoritmos baseados no *kernel* gaussiano assumem que os exemplos de conjuntos de dados são mais prováveis de estarem distribuídos dentro de uma região hiper-esférica (em outras palavras, cada variável tem a mesma variância e não há correlação entre variáveis). Entretanto, exemplos em dois grupos diferentes são mais prováveis de estarem distribuídos dentro de duas regiões hiper-elipsoidais diferentes.

A distância de Mahalanobis, que leva em consideração as correlações entre variáveis e é invariante em escala, é uma melhor escolha para lidar com regiões hiper-elipsoidais (CAMPS-VALLS *et al.* [22], ABE [23], KAMADA e ABE [24], WANG *et al.* [25]). FERREIRA e DE CARVALHO [6] propuseram, sob o enfoque de kernelização da métrica, um algoritmo do tipo *kernel K-médias* baseado em um *kernel* de Mahalanobis adaptativo. Este *kernel* foi construído a partir de uma distância quadrática adaptativa definida por uma matriz simétrica positiva-definida que é modificada a cada iteração do algoritmo. Contudo, o método proposto por FERREIRA e DE CARVALHO [6] considera apenas o caso em que a matriz de covariâncias é a mesma para todos os grupos, tendo a matriz diagonal (caso em que não há covariâncias entre as variáveis) como caso particular, ou seja, os grupos podem assumir formas hiper-elipsoidais mas há a imposição de que essas regiões sejam iguais. Nesta dissertação apresentaremos a versão *fuzzy* do método proposto por (FERREIRA e DE CARVALHO [6]) e suas extensões para o caso em que a matriz de covariâncias pode ser diferente para cada grupo, isto é, os grupos podem assumir formas hiper-elipsoidais e essas regiões podem ser diferentes entre os grupos. Dessa forma, uma classe muito maior de problemas de agrupamento poderá ser abordada.

1.1 Objetivos

Os objetivos deste trabalho são:

- Propor métodos de agrupamento fuzzy baseado no Kernel de Mahalanobis com distâncias quadráticas adaptativas e apresentar os casos particulares para diferentes matrizes de covariância;
- Dissertar sobre os métodos K-mean e Fuzzy K-mean como embasamento teórico para essa dissertação, além de apresentar alguns casos particulares;
- Programar o código dos métodos de agrupamento fuzzy baseados nas diferentes matrizes de covariância;
- Fazer uma esquema experimental com conjuntos de dados numéricos e reais, avaliando a eficiência dos algoritmos propostos.

Capítulo 2

Revisão Bibliográfica

Dentre as técnicas de análise multivariada a *análise de agrupamento* é uma das mais populares sendo utilizada nas mais diversas áreas. Segundo MANLY [3] agrupar uma classe de valores por meio de características similares se constitui um esquema para definir grupos e estabelecer padrões entre as classificações das variáveis. A análise de agrupamento pode ser aplicada para:

- Dividir os elementos amostrais em grupos de modo que haja similaridade entre os elementos que pertencem ao mesmo grupo e haja dissimilaridade entre elementos que pertencem a grupos distintos;
- Dividir as variáveis observadas de modo que aquelas que pertençam a um mesmo grupo tenham alta correlação entre si e baixa correlação com as variáveis pertencentes aos outros grupos.

Nesta seção descreveremos duas abordagens de análise de agrupamentos: as *Hierárquicas* e as *Não - Hierárquicas ou Particionais*. As *Hierárquicas* são métodos simples que utilizam os dados tanto para produzir uma partição para um número de grupos pré-fixado a priori denotado por c , quanto para estimar o número de grupos. As *Não - Hierárquicas ou Particionais* são métodos baseados na minimização de uma função objetivo, onde os padrões são agrupados em um número c de grupos escolhido a priori. Cada padrão é agrupado na classe em que essa função objetivo é minimizada. No conjunto das técnicas não-hierárquicas tem-se os métodos hard, que produzem partições que não permitem interseção entre grupos, onde cada elemento do banco de dados é alocado em um único grupo. E os métodos fuzzy, que são técnicas de agrupamento de dados onde cada padrão pertence a um grupo com um certo grau de pertinência.

Os métodos de agrupamentos *hierárquicos* são fundamentados em funções de ligação que são usadas em cada passo do algoritmo para comparação dos grupos disponíveis para agrupamentos. Existem várias funções de ligação e cada uma gera

um método de agrupamento diferente (MINGOTI [4]). Alguns das funções de ligação mais tradicionais são visto na Tabela 2.1:

Tabela 2.1: Funções de Ligação

Método	Fórmula
Ligação Simples	$d(i, j) = \min d(i, j)$, em que $i \in C_1, j \in C_2$
Ligação Completa	$d(i, j) = \max d(i, j)$, em que $i \in C_1, j \in C_2$
Ligação Média	$d(i, j) = \frac{1}{n_i, n_j} \sum_{i, j \in C_1, C_2} d(i, j)$, em que $i \in C_1, j \in C_2$
Centroídes	$d(i, j) = d(\mu_i, \mu_j)$, em que $\mu_i \in C_1, \mu_j \in C_2$

O Método de Ligação Simples se baseia na distância mínima entre as observações e os grupos formados, em que usamos a matriz original de distância entre observações nos dois grupos para escolher a menor distância como medida de similaridade entre os grupos. O Método da Ligação Completa é parecido com o método de ligação simples, em que a similaridade se baseia na distância máxima entre observações em cada grupo. No Método de Ligação Média sua similaridade é baseada em todos os elementos dos agregados, e não em um único par de membros externos, e é desse modo menos afetada por observações atípicas (MANLY [3]).

Independente da função de ligação utilizada para construção do agrupamento, os grupos que devem ser unidos no respectivo passo do algoritmo são aqueles de maior ou menor similaridade. A similaridade entre dois indivíduos é medida através de um coeficiente que representa a quantificação da semelhança entre duas linhas (indivíduos) da matriz de dados. Dentre os vários tipos de coeficientes que podem ser utilizados, merecem destaque especial as distâncias e a correlação. Existem outros métodos que se utilizam de medidas de tendências central para sintetizar as informações dos grupos que vão sendo construídos em cada passo. A formação de novos grupos é feita com base nas distâncias entre grupos calculadas usando a medida de tendência central escolhida como representante das informações dos grupos, um desses métodos mais conhecido é o centróide.

No Método do Centróide, a similaridade entre dois agrupamentos é a distância entre os vetores de médias dos grupos. Centróides são os valores médios das observações sobre as variáveis na variável estatística de agrupamento. Neste método, toda vez que indivíduos são reunidos, um novo centróide é computado (MANLY [3]). Isso ocorre porque existe mudança no centróides dos agrupamentos toda vez que um novo indivíduo ou grupo de indivíduos é acrescentado a um grupo já existente. Logo, considerando dois grupos C_1 e C_2 e seus respectivos vetores de médias amostrais $\bar{\mathbf{X}}_1$ e $\bar{\mathbf{X}}_2$, a distância entre C_1 e C_2 é definida por:

$$d(C_1, C_2) = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^\top (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \quad (2.1)$$

que é a distância Euclidiana ao quadrado entre os vetores de médias amostral \bar{X}_1 e \bar{X}_2 . O método do centroide também pode ser usado com a distância Euclidiana usual entre os vetores de médias. Em cada passo do algoritmo do agrupamento, os conglomerados que apresentam o menor valor de distância são agrupados (MINGOTI [4]).

2.1 Técnicas Não-Hierárquicas

As técnicas hierárquicas aglomerativas são boas ferramentas para produzir uma partição inicial dos dados. No entanto, a propriedade de hierarquia causa o problema de não dissociação de grupos que foram unidos em determinado passo do algoritmo já que somente é permitido a entrada de elementos nos grupos já formados. As técnicas não-hierárquicas também chamadas de *métodos de partição*, por sua vez, são mais flexíveis e permitem que elementos amostrais mudem de grupo em cada passo do algoritmo se essa mudança permitir uma melhor solução em termos de variabilidade da partição resultante. O objetivo básico dessas técnicas é buscar diretamente uma partição dos dados em c grupos de modo a preservar coesão interna e isolamento entre elementos. Dentre as técnicas não - hierárquicas apresentaremos nessa seção o método c -Média e o Fuzzy c -Média.

2.1.1 Método c -Médias

No método c -Médias os grupos são definidos a priori, em que cada elemento é alocado em apenas um único grupo dentre os c disponíveis. Calcula-se os pontos dos centroídes heurísticamente para cada grupo até alcançar um equilíbrio estático nos conjuntos de pontos para cada grupo MACQUEEN [26]. O método c -Média busca a partição dos dados em c grupos que minimiza a seguinte função objetivo:

$$J = \sum_{k=1}^n \sum_{i=1}^c (u_{ik}) \phi(\mathbf{x}_k, \mathbf{g}_i) = \sum_{k=1}^n \sum_{i=1}^c (u_{ik}) \sum_{j=1}^p (\mathbf{x}_{kj} - \mathbf{g}_{ij})^2, \quad (2.2)$$

Em que

$$\phi(\mathbf{x}_k, \mathbf{g}_i) = \sum_{j=1}^p (\mathbf{x}_{kj} - \mathbf{g}_{ij})^2, \quad (2.3)$$

é o quadrado da distância euclidiana que mede a dissimilaridade entre um par de vetores de valores de característica quantitativa, $\mathbf{x}_k = (\mathbf{x}_{k1}, \dots, \mathbf{x}_{kp})$ é o vetor de

características quantitativas que descreve a k -ésima observação, $\mathbf{g}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{ip})$ são os centróides dos grupos.

O algoritmo depende de um parâmetro (c =número de grupos) definido de forma a priori pelo pesquisador. Isto costuma ser um problema, tendo em vista que normalmente não se sabe quantos grupos existem a priori, pois em muitas situações tem-se elementos amostrais com características semelhantes entre mais de um grupo e que eventualmente não estarão bem alocados na partição obtida pelo método c -média. Este algoritmo é extremamente veloz, geralmente convergindo em poucas iterações para uma configuração estável, na qual nenhum elemento está designado para um grupo cujo centro não lhe seja o mais próximo. O algoritmo c -médias é executado de acordo com os seguintes passos:

1. Inicialização

Partição inicial dos sujeitos em k grupos definidos à priori;

2. Etapa 1: Cálculo dos centróides para cada um dos k grupos

3. Etapa 2 : Cálculo da distância euclidiana dos centróides a cada indivíduo na base de dados;

4. Agrupar os indivíduos aos grupos de cujos centróides se encontram mais próximos e voltar ao passo 2 até que não ocorra uma variação significativa na distância mínima de cada indivíduo da base de dados a cada um dos centróides dos c grupos.

2.1.2 Método *Fuzzy* c -Médias

O método *Fuzzy* c -Médias (BEZDEK [27]) tem como objetivo estimar um coeficiente chamado de *grau de pertinência*, para cada elemento do banco de dados em relação a cada um dos c grupos. Elementos que apresentam graus de pertinência similares para dois ou mais grupos são elementos fuzzy de fronteiras e são as causas dos prováveis problemas de má alocação em procedimentos como o c -Médias, por exemplo. O grau de pertinência pode ser interpretado como uma medida de chance de que o elemento faça parte do respectivo grupo.

O método *Fuzzy* c -Médias se inicia com a escolha de c sementes chamadas 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. Seja Ω um conjunto não vazio. Um conjunto fuzzy em Ω é uma função

$$X : \Omega \mapsto [0, 1]$$

$$n \mapsto X(n).$$

O algoritmo de FCM (Fuzzy C - Médias) é um método de agrupamento não hierárquico cujo objetivo é fornecer uma partição fuzzy de um conjunto de padrões em c grupos $\{\mathbf{P}_1, \dots, \mathbf{P}_c\}$ e um conjunto correspondente de centróides $\{\mathbf{g}_1, \dots, \mathbf{g}_c\}$, de modo que um critério J que mede o ajuste entre os grupos e seus centróides são minimizados localmente. Este critério J é definido como:

$$J = \sum_{k=1}^n \sum_{i=1}^c (u_{ik})^m \phi(\mathbf{x}_k, \mathbf{g}_i) = \sum_{k=1}^n \sum_{i=1}^c (u_{ik})^m \sum_{j=1}^p (\mathbf{x}_{kj} - \mathbf{g}_{ij})^2, \quad (2.4)$$

Em que

$$\phi(\mathbf{x}_k, \mathbf{g}_i) = \sum_{j=1}^p (\mathbf{x}_{kj} - \mathbf{g}_{ij})^2, \quad (2.5)$$

é o quadrado da distância euclidiana que mede a dissimilaridade entre um par de vetores de valores de característica quantitativa, $\mathbf{x}_k = (\mathbf{x}_{k1}, \dots, \mathbf{x}_{kp})$ é o vetor de características quantitativas que descreve a k -ésima observação, $\mathbf{g}_i = (\mathbf{g}_{i1}, \dots, \mathbf{g}_{ip})$ são os centróides dos grupos $P_i, i = 1, \dots, c$, u_{ik} é o grau de pertinência da k -ésima observação para o i -ésimo grupo P_i e $m \in (1, \infty)$ é o parâmetro que controla a variação da associação para cada padrão k .

Na etapa de representação, o grau de associação de cada padrão K no grupo P_i é fixo e o centróide $\mathbf{g}_i = (\mathbf{g}_{i1}, \dots, \mathbf{g}_{ip})$ do grupo P_i ($i = 1, \dots, c$) que minimiza o critério de agrupamento J , é atualizado de acordo com a seguinte expressão:

$$\mathbf{g}_i = \frac{\sum_{k=1}^n (u_{ik})^m \mathbf{x}_k}{\sum_{k=1}^n (u_{ik})^m}. \quad (2.6)$$

Na etapa de alocação, cada centróide \mathbf{g}_i do grupo P_i ($i = 1, \dots, c$) é fixo e os graus de pertinência u_{ik} ($k = 1, \dots, n$) de cada padrão k em cada grupo P_i , minimizam o agrupamento do critério J sob $u_{ik} \geq 0$ e $\sum_{i=1}^c u_{ik} = 1$, são atualizados de acordo com a seguinte expressão:

$$u_{ik} = \left[\sum_{h=1}^c \left\{ \frac{\sum_{j=1}^p (\mathbf{x}_{kj} - \mathbf{g}_{ij})^2}{\sum_{j=1}^p (\mathbf{x}_{kj} - \mathbf{g}_{hj})^2} \right\}^{\frac{1}{m-1}} \right]^{-1} \quad \text{para } i = 1, \dots, c. \quad (2.7)$$

O algoritmo fuzzy c -médias 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 (um número máximo de iterações); e fixe $0 < \varepsilon \ll 1$; inicialize os graus de pertinência difusos u_{ik} ($k = 1, \dots, c, i = 1, \dots, n$) de tal forma que $u_{ik} \geq 0 \forall i, k$ e $\sum_{k=1}^c u_{ik} = 1 \forall i$, faça $t = 1$.

2. Etapa 1: Definição dos melhores centroides dos grupos. Atualize os centroides dos grupos \mathbf{g}_k ($k = 1, \dots, c$) de acordo com a Eq (2.6).
3. Etapa 2: Definição da melhor partição difusa Atualize os graus de pertinências difusos u_{ik} ($k = 1, \dots, c, i = 1, \dots, n$) de acordo com a Eq. (2.7)
4. Critério de parada Se $|J^{t+1} - J^t| \leq \varepsilon$ ou $t > T$, então, *PARE*, caso contrário, faça $t = t + 1$ e volte ao passo (2).

2.1.3 Fuzzy c-Médias baseado em Kernel

Nesta seção apresentaremos uma breve revisão acerca da teoria básica sobre funções *kernel* e sobre as versões convencionais do algoritmos *kernel c-médias*.

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 (MERCER [28]):

$$\sum_{i=1}^n \sum_{k=1}^n c_i c_k \mathcal{K}(\mathbf{x}_i, \mathbf{x}_k) \geq 0 \quad \forall n \geq 2, \quad (2.8)$$

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

Um conjunto de padrões de entrada não-linearmente separável pode tornar-se separável linearmente através de um mapeamento não-linear arbitrário para um espaço de características de alta dimensão (HAYKIN [29]). Seja $\Phi : X \rightarrow \mathcal{F}$ um mapeamento não-linear arbitrário do espaço original X para um espaço de características de alta dimensão \mathcal{F} . Aplicando o mapeamento não-linear Φ , o produto interno $\mathbf{x}_i^\top \mathbf{x}_k$ no espaço original é mapeado para $\Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_k)$ no espaço de características. A essência 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

$$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_k) = \Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_k), \quad (2.9)$$

que é usualmente referida como *kernel trick* (SCHÖLKOPF *et al.* [30], MÜLLER *et al.* [31]).

Seja $\mathbf{K} \in \mathbb{R}^{n \times n}$ uma matriz chamada de matriz *kernel* onde cada elemento $\kappa_{il} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_l)$, $i = 1, \dots, n$, $l = 1, \dots, n$.

De Eq. (2.9), é possível calcular distâncias Euclidianas em \mathcal{F} da seguinte maneira (SCHÖLKOPF *et al.* [30], MÜLLER *et al.* [31]):

$$\begin{aligned} \|\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_k)\|^2 &= (\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_k))^\top (\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_k)) \\ &= \Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_i) - 2\Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_k) + \Phi(\mathbf{x}_k)^\top \Phi(\mathbf{x}_k) \\ &= \mathcal{K}(\mathbf{x}_i, \mathbf{x}_i) - 2\mathcal{K}(\mathbf{x}_i, \mathbf{x}_k) + \mathcal{K}(\mathbf{x}_k, \mathbf{x}_k). \end{aligned} \quad (2.10)$$

Exemplos de funções *kernel* tipicamente utilizadas são:

- 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: $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_k) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_k\|^2}{2\sigma^2}}$, $\sigma > 0$,
- Laplaciana: $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_k) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_k\|}$, $\gamma > 0$,
- Sigmóide: $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_k) = \tanh(\gamma \mathbf{x}_i^\top \mathbf{x}_k + \theta)$, $\gamma > 0$, $\theta \geq 0$,

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

Existem duas grandes variações dos métodos de agrupamento baseados em *kernel*, as quais são baseadas, respectivamente, em: kernelização da métrica e agrupamento no espaço de características. Métodos de agrupamento baseados em kernelização da métrica buscam por centroídes no espaço original dos padrões e a distância entre um padrão \mathbf{x}_i e o centroíde do k -ésimo grupo \mathbf{g}_k é calculada por meio de funções *kernel*: $\|\Phi(\mathbf{x}_i) - \Phi(\mathbf{g}_k)\|^2 = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_i) - 2\mathcal{K}(\mathbf{x}_i, \mathbf{g}_k) + \mathcal{K}(\mathbf{g}_k, \mathbf{g}_k)$.

Por outro lado, algoritmos de agrupamento no espaço de características realizam um mapeamento de cada padrão por meio de uma função não-linear Φ e então obtêm os centroídes dos grupos no espaço de características. Seja \mathbf{g}_k^Φ o centroíde do k -ésimo grupo no espaço de características. Nós veremos que é possível calcular $\|\Phi(\mathbf{x}_i) - \mathbf{g}_k^\Phi\|^2$ sem a necessidade de se obter \mathbf{g}_k^Φ , através do uso do *kernel trick* (Eq. (2.9)). A principal vantagem é que, ao passarmos para um espaço de mais alta dimensão, um conjunto de padrões de entrada não-linearmente separável pode tornar-se separável linearmente, e, embora não conheçamos o mapeamento não-linear Φ e não possamos obter os centroídes dos grupos, as distâncias entre os padrões e os centroídes dos grupos podem ser calculadas através de funções *kernel*.

2.1.4 Kernel c -médias difuso baseado em Kernelização da métrica

A idéia no algoritmo Kernel c -médias difuso baseado em kernelização da métrica é minimizar a seguinte função objetivo:

$$\begin{aligned} J &= \sum_{k=1}^c \sum_{i=1}^n (\mathbf{u}_{ik})^m \|\Phi(\mathbf{x}_i) - \Phi(\mathbf{g}_k)\|^2 \\ &= \sum_{k=1}^c \sum_{i=1}^n (\mathbf{u}_{ik})^m \{K(\mathbf{x}_i, \mathbf{x}_i) - 2K(\mathbf{x}_i, \mathbf{g}_k) + K(\mathbf{g}_k, \mathbf{g}_k)\}, \end{aligned} \quad (2.11)$$

sob as restrições

$$\begin{cases} \mathbf{u}_{ik} \in [0, 1] & \forall k, i \\ \sum_{k=1}^c \mathbf{u}_{ik} = 1 & \forall i, \end{cases} \quad (2.12)$$

onde $\mathbf{g}_k \in \mathbb{R}^p$ é o centroíde do k -ésimo grupo ($k = 1, \dots, c$), u_{ik} é o grau de pertinência difuso do padrão i ao k -ésimo grupo ($k = 1, \dots, K, i = 1, \dots, n$) e $m \in (1, \infty)$ é um parâmetro que controla o grau de imprecisão da pertinência para cada padrão i . Seja $U = (u_{ik}), k = 1, \dots, c, i = 1, \dots, n$. U é chamada de matriz de partição difusa. Como no caso do algoritmo *kernel* K -médias rígido baseado em kernelização da métrica, a derivação dos protótipos dos grupos depende da escolha da função *kernel*. Considerando o *kernel* Gaussiano, a função objetivo dada pela Eq. (2.11) fica reescrita da forma:

$$J = 2 \sum_{k=1}^c \sum_{i=1}^n (u_{ik})^m (1 - K(\mathbf{x}_i, \mathbf{g}_k)). \quad (2.13)$$

Assim, a equação de atualização dos protótipos dos grupos fica expressa como:

$$\mathbf{g}_k = \frac{\sum_{i=1}^n (u_{ik})^m K(\mathbf{x}_i, \mathbf{g}_k) \mathbf{x}_i}{\sum_{i=1}^n (u_{ik})^m K(\mathbf{x}_i, \mathbf{g}_k)}, \quad (2.14)$$

Na etapa de atualização da matriz de partição difusa U , os centroídes dos grupos \mathbf{g}_k ($k = 1, \dots, c$) são mantidos fixos. Precisamos então encontrar os graus de pertinência difusos u_{ik} ($k = 1, \dots, c, i = 1, \dots, n$) que minimizam a função objetivo J sob as restrições dadas em (2.12). Através do método dos multiplicadores de Lagrange, chegamos à seguinte solução:

$$u_{ik} = \left[\sum_{h=1}^c \left(\frac{1 - K(\mathbf{x}_i, \mathbf{g}_k)}{1 - K(\mathbf{x}_i, \mathbf{g}_h)} \right)^{\frac{1}{m-1}} \right]^{-1}. \quad (2.15)$$

O algoritmo *kernel c*-médias difuso baseado em kernelização da métrica 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 (um número máximo de iterações); e fixe $0 < \varepsilon \ll 1$; inicialize os graus de pertinência difusos u_{ik} ($k = 1, \dots, c, i = 1, \dots, n$) de tal forma que $u_{ik} \geq 0 \forall i, k$ e $\sum_{k=1}^c u_{ik} = 1 \forall i$, faça $t = 1$.

2. Etapa 1: Definição dos melhores centroídes dos grupos. Atualize os centroídes dos grupos \mathbf{g}_k ($k = 1, \dots, c$) de acordo com a Eq (2.14).

3. Etapa 2: Definição da melhor partição difusa

Atualize os graus de pertinências difusos u_{ik} ($k = 1, \dots, c, i = 1, \dots, n$) de acordo com a Eq. (2.15)

4. Critério de parada

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

2.1.5 Kernel *c*-médias difuso no espaço de características

O algoritmo *Kernel c*-médias difuso no espaço de características busca iterativamente por c grupos através da minimização da seguinte função objetivo:

$$J = \sum_{k=1}^c \sum_{i=1}^n (u_{ik})^m \|\Phi(\mathbf{x}_i) - (\mathbf{g}_k^\Phi)\|^2, \quad (2.16)$$

sujeita às restrições dadas na Eq. (2.12), onde u_{ik} ($k = 1, \dots, c, i = 1, \dots, n$) e m são definidos como anteriormente e \mathbf{g}_k^Φ é o centroíde do k -ésimo grupo no espaço de características.

A minimização da função objetivo dada pela Eq. (2.16) com respeito a \mathbf{g}_k^Φ fornece a seguinte expressão de atualização dos centroídes dos grupos no espaço de características:

$$\mathbf{g}_k^\Phi = \frac{\sum_{i=1}^n (u_{ik})^m \Phi(\mathbf{x}_i)}{\sum_{i=1}^n (u_{ik})^m}, \quad k = 1, \dots, c. \quad (2.17)$$

O próximo passo é minimizar a Eq. (2.16) com respeito a u_{ik} ($k = 1, \dots, c, i = 1, \dots, n$) sob as restrições dadas em (2.12). Novamente, a utilização do método dos multiplicadores de Lagrange leva à seguinte solução:

$$u_{ik} = \left[\sum_{h=1}^c \left(\frac{\|\Phi(\mathbf{x}_i) - \mathbf{g}_k^\Phi\|^2}{\|\Phi(\mathbf{x}_i) - \mathbf{g}_h^\Phi\|^2} \right)^{\frac{1}{m-1}} \right]^{-1}. \quad (2.18)$$

Como no caso rígido, a distância entre $\Phi(\mathbf{x}_i)$ e \mathbf{g}_h^Φ no espaço de característica é calculada através de funções *kernel* no espaço original dos padrões:

$$\begin{aligned} \|\Phi(\mathbf{x}_i) - \mathbf{g}_k^\Phi\|^2 &= \Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_i) - 2\Phi(\mathbf{x}_i)^\top (\mathbf{g}_k^\Phi) + (\mathbf{g}_k^\Phi)^\top (\mathbf{g}_k^\Phi) \\ &= \Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_i) - \frac{2 \sum_{l=1}^n (u_{lk})^m \Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_l)}{\sum_{l=1}^n (u_{lk})^m} \\ &\quad + \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{rk})^m (u_{sk})^m \Phi(\mathbf{x}_r)^\top \Phi(\mathbf{x}_s)}{(\sum_{r=1}^n (u_{rk})^m)^2} = \\ &= K(x_i, x_i) - \frac{2 \sum_{l=1}^n (u_{lk})^m K(x_l, x_i)}{\sum_{l=1}^n (u_{lk})^m} \\ &\quad + \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{rk})^m (u_{sk})^m K(x_r, x_s)}{(\sum_{r=1}^n ((u_{rk})^m))^2}. \end{aligned} \quad (2.19)$$

Adicionalmente, a função objetivo J dada pela Eq. 2.16 fica reescrita da forma:

$$\begin{aligned} J &= \sum_{k=1}^c \sum_{i=1}^n (u_{ik})^m \left\{ K(\mathbf{x}_i, \mathbf{x}_i) - \frac{2 \sum_{l=1}^n (u_{lk})^m K(\mathbf{x}_l, \mathbf{x}_i)}{\sum_{l=1}^n (u_{lk})^m} \right. \\ &\quad \left. + \frac{\sum_{r=1}^n \sum_{s=1}^n (u_{rk})^m (u_{sk})^m K(\mathbf{x}_r, \mathbf{x}_s)}{(\sum_{r=1}^n (u_{rk})^m)^2} \right\} \end{aligned} \quad (2.20)$$

O algoritmo *kernel* c -médias difuso no espaço de características também não possui a etapa em que os protótipos dos grupos são atualizados. A atualização da matriz de partição difusa U pode ser feita sem o cálculo dos centroídes devido ao mapeamento implícito realizado através na função *kernel* na Eq.(2.19).

O algoritmo *kernel* c -médias 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 (um número máximo de iterações); e fixe $0 < \varepsilon \ll 1$; inicialize os graus de pertinência difusos u_{ik} ($k = 1, \dots, c, i = 1, \dots, n$) de tal forma que $u_{ik} \geq 0 \forall i, k$ e $\sum_{k=1}^c u_{ik} = 1 \forall i$; escolha uma função *kernel* adequada k e calcule a matriz *kernel* \mathbf{K} ; faça $t = 1$.

2. Etapa 1: Definição da melhor partição difusa

Atualize os graus de pertinências difusos u_{ik} ($k = 1, \dots, c, i = 1, \dots, n$) de acordo com a Eq. (2.18).

3. Critério de parada

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

Capítulo 3

Método Proposto

3.1 Introdução

Neste capítulo encontra-se a principal contribuição desta dissertação de mestrado. São introduzidos métodos de agrupamento fuzzy baseados no Kernel de Mahalanobis com distâncias quadráticas adaptativas. Tal distância foi definida através de matrizes de covariâncias fuzzy completa ou diagonal, podendo as matrizes serem única ou diferente para cada grupo.

A ideia principal do método de agrupamento fuzzy baseado em uma única distância adaptativa consiste em comparar grupos e seus centróides por meio de uma única distância que muda em cada iteração do algoritmo, mas é a mesma para todos os grupos. No caso do método de agrupamento fuzzy baseado em distância adaptativa diferente para cada grupo, compara-se os grupos e centróides usando uma distância diferente associada a cada grupo que muda em cada iteração do algoritmo, ou seja, a distância não é determinada de forma absoluta e é diferente de um grupo para o outro.

Esses métodos adaptativos procuram uma partição fuzzy de um conjunto de padrões em c grupos que corresponde a um conjunto de c centróides $\{g_1, \dots, g_c\}$ e uma distância quadrática adaptativa única ou diferente para cada grupo, de modo que um critério que mede o ajuste entre os grupos e seus representantes (centróides) seja minimizado localmente.

O *kernel* Gaussiano é o mais comumente utilizado na literatura. Em geral, este *kernel* fornece bons resultados e requer somente o ajuste de um parâmetro. O *kernel* Gaussiano é baseado na distância euclidiana entre dois padrões \mathbf{x}_i e \mathbf{x}_k em \mathbb{R}^p e se sabe que a distância euclidiana fornece bons resultados quando os grupos são aproximadamente hiperesféricos, o que significa que cada variável tem a mesma variância

e que não há correlação entre elas. Contudo, observações em dois diferentes grupos são mais prováveis de estarem distribuídas dentro de duas regiões hiperelipsoidais diferentes. A distância de Mahalanobis, que leva em conta as covariâncias entre as variáveis e é invariante sob mudança de escala, é, desse modo, uma escolha mais apropriada para lidar com grupos de formas hiperelipsoidais, as quais surgem em diversas situações práticas e experimentais.

A distância de Mahalanobis entre uma observação \mathbf{x}_i e o centróide global \mathbf{g} em \mathbb{R}^p é dada por

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

em que

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

Inspirados pela definição da distância de Mahalanobis, podemos definir um *kernel* adaptativo de Mahalanobis substituindo a distância euclidiana no *kernel* Gaussiano por uma distância tipo-Mahalanobis (CAMPS-VALLS *et al.* [22]):

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

onde $d_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_k) = (\mathbf{x}_i - \mathbf{x}_k)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{x}_k)$ é uma distância quadrática definida por uma matriz simétrica positiva-definida \mathbf{M} . Aqui, a distância de Mahalanobis é calculada entre duas observações \mathbf{x}_i e \mathbf{x}_k em \mathbb{R}^p ao invés de uma observação \mathbf{x}_i e o centróide global \mathbf{y} . O *kernel* de Mahalanobis é uma extensão do Gaussiano. Note que, fazendo $\mathbf{M} = \mathbf{I}$, onde \mathbf{I} é a matriz identidade de dimensão $p \times p$, temos o *kernel* Gaussiano.

Baseados no conceito de distância de Mahalanobis, nesta dissertação iremos considerar quatro distâncias quadráticas adaptativas entre uma observação \mathbf{x}_i e o centróide do k -ésimo grupo \mathbf{y}_k definidos como:

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

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

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

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

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

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

(d) Distância quadrática adaptativa definida por uma matriz de covariâncias diagonal para cada grupo:

$$\begin{aligned}
\mathbf{M}_k &= \text{diag}\{\lambda_1^k, \dots, \lambda_p^k\} : d_{\mathbf{M}_k}^2(\mathbf{x}_i, \mathbf{g}_k) = (\mathbf{x}_i - \mathbf{g}_k)^\top \mathbf{M}_k(\mathbf{x}_i - \mathbf{g}_k) \\
&= \sum_{j=1}^p \lambda_j^k (\mathbf{x}_i - \mathbf{g}_k)^2, k = 1, \dots, c.
\end{aligned} \tag{3.6}$$

3.2 Método de agrupamento fuzzy baseado no Kernel de Mahalanobis com distâncias quadráticas adaptativas

Nesta seção apresentaremos os métodos de agrupamentos fuzzy baseados no kernel de Mahalanobis com distâncias quadráticas adaptativas. Os algoritmos introduzidos nesta seção otimizam uma função objetivo J que mede o ajuste entre os grupos e os seus centróides, a qual pode ser definida genericamente como

$$J = \sum_{i=1}^n \sum_{k=1}^c (u_{ik})^m \varphi^2(\mathbf{x}_i, \mathbf{g}_k) \tag{3.7}$$

sujeita às restrições dadas na Eq. (2.12), onde $\varphi^2(\mathbf{x}_i, \mathbf{g}_k)$ é uma medida de distância adequada entre um padrão \mathbf{x}_i e o centróide do k -ésimo grupo \mathbf{g}_k calculada por meio de uma função kernel de Mahalanobis, u_{ik} , ($k = 1, \dots, c, i = 1, \dots, n$) e m são definidos como na Seção 2.1.3. Considerando as medidas de distâncias propostas

neste capítulo, temos:

(a) Método de agrupamento fuzzy baseado no Kernel de Mahalanobis com distância quadrática adaptativa definida por uma matriz de covariância fuzzy completa (MK-FCM-PFCV):

$$\begin{aligned}
J1 &= \sum_{i=1}^n \sum_{k=1}^c (u_{ik})^m \|\phi(\mathbf{x}_i) - \phi(\mathbf{g}_k)\|^2 \\
&= 2 \sum_{i=1}^n \sum_{k=1}^c (u_{ik})^m \left\{ 1 - \exp \left\{ -\frac{(\mathbf{x}_i - \mathbf{g}_k)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{g}_k)}{2\sigma^2} \right\} \right\}. \quad (3.8)
\end{aligned}$$

onde \mathbf{x}_i , \mathbf{g}_k , u_{ik} e \mathbf{M} são definidos como antes e,

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

onde

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

é uma distância quadrática definida por uma matriz simétrica definida positiva \mathbf{M} .

(b) Método de agrupamento fuzzy baseado no Kernel de Mahalanobis com distância quadrática adaptativa definida por uma matriz de covariância fuzzy diagonal (MK-FCM-PDFCV):

$$\begin{aligned}
J2 &= \sum_{i=1}^n \sum_{k=1}^c (u_{ik})^m \|\phi(\mathbf{x}_i) - \phi(\mathbf{g}_k)\|^2 \\
&= 2 \sum_{i=1}^n \sum_{k=1}^c (u_{ik})^m \left\{ 1 - \exp \left\{ -\frac{(\mathbf{x}_i - \mathbf{g}_k)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{g}_k)}{2\sigma^2} \right\} \right\}. \quad (3.9)
\end{aligned}$$

onde \mathbf{x}_i , \mathbf{g}_k , u_{ik} e \mathbf{M} são definidos como antes e,

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

e a matriz \mathbf{M} é definida como:

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

onde $d_M^2(\mathbf{x}_i, \mathbf{g}_k)$ 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 baseado no Kernel de Mahalanobis com distâncias quadráticas adaptativas diferentes para cada grupo definidas por uma matriz de covariância fuzzy completa (MK-FCM-FCV):

$$\begin{aligned} J3 &= \sum_{i=1}^n \sum_{k=1}^c (u_{ik})^m \|\phi(\mathbf{x}_i) - \phi(\mathbf{g}_k)\|^2 \\ &= 2 \sum_{i=1}^n \sum_{k=1}^c (u_{ik})^m \left\{ 1 - \exp \left\{ -\frac{(\mathbf{x}_i - \mathbf{g}_k)^\top \mathbf{M}_k (\mathbf{x}_i - \mathbf{g}_k)}{2\sigma^2} \right\} \right\}. \end{aligned} \quad (3.10)$$

onde $\mathbf{x}_i, \mathbf{g}_k, u_{ik}$ são definidos como antes e,

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

onde

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

é uma distância quadrática definida por uma matriz simétrica definida positiva \mathbf{M}_k associada a cada grupo i ($i = 1, \dots, c$).

(d) Método de agrupamento fuzzy baseado no Kernel de Mahalanobis com distâncias quadráticas adaptativas diferentes para cada grupo definidas por uma matriz de covariância fuzzy diagonal (MK-FCM-DFCV):

$$\begin{aligned} J4 &= \sum_{i=1}^n \sum_{k=1}^c (u_{ik})^m \|\phi(\mathbf{x}_i) - \phi(\mathbf{g}_k)\|^2 \\ &= 2 \sum_{i=1}^n \sum_{k=1}^c (u_{ik})^m \left\{ 1 - \exp \left\{ -\frac{(\mathbf{x}_i - \mathbf{g}_k)^\top \mathbf{M}_k (\mathbf{x}_i - \mathbf{g}_k)}{2\sigma^2} \right\} \right\}. \end{aligned} \quad (3.11)$$

onde $\mathbf{x}_i, \mathbf{g}_k, u_{ik}$ e \mathbf{M}_k são definidos como antes e,

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

e a matriz \mathbf{M}_k é definida como:

$$\mathbf{M}_k = \begin{pmatrix} \lambda_{1k} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_{ck} \end{pmatrix}$$

onde $\mathbf{d}_M^2(\mathbf{x}_i, \mathbf{g}_k)$ é uma distância quadrática definida por uma matriz simétrica diagonal definida positiva \mathbf{M} que é diferente para todos os grupos.

Para organização e desenvolvimento do método de agrupamento fuzzy baseado no Kernel de Mahalanobis formularemos o algoritmo em três etapas. A primeira etapa consiste em calcular os centróides. Na derivação dos centróides dos grupos, os graus de pertinência fuzzy u_{ik} ($k = 1, \dots, c, i = 1, \dots, n$) e as matrizes de covariância fuzzy são mantidos fixos.

Proposição 3.2.1. *Qualquer que seja a medida de distância quadrática adaptativa adotada Eqs. (3.3), (3.4), (3.5), (3.6), e se $\mathcal{K}^{(m)}(.,.)$ é o Kernel de Mahalanobis, então, o centróide $\mathbf{g}_k = (g_1, \dots, g_c)$ ($k = 1, \dots, c$) que minimiza a função objetivo J , dada pela Eq. (3.7), tem seus componentes g_k ($k = 1, \dots, c$) atualizados de acordo com a seguinte expressão:*

$$g_k = \frac{\sum_{i=1}^n (u_{ik})^m \mathcal{K}^{(m)}(\mathbf{x}_i, \mathbf{g}_k) \mathbf{x}_i}{\sum_{i=1}^n (u_{ik})^m \mathcal{K}^{(m)}(\mathbf{x}_i, \mathbf{g}_k)}$$

Demonstração: Se nos restringirmos ao Kernel de Mahalanobis, então, $\mathcal{K}_j^{(m)}(\mathbf{x}_i, \mathbf{x}_i) = 1$, ($i = 1, \dots, n$) e $\|\Phi(\mathbf{x}_i) - \Phi(\mathbf{g}_k)\|^2 = 2\{1 - \mathcal{K}^{(m)}(\mathbf{x}_i, \mathbf{g}_k)\}$. Dessa forma, para o k -ésimo grupo o problema se torna encontrar o componente g_k que maximiza o termo:

$$\sum_{i=1}^n (u_{ik})^m \mathcal{K}^{(m)}(\mathbf{x}_i, \mathbf{g}_k) \quad (3.12)$$

Tomando a primeira derivada do termo dado em 3.12 com relação a \mathbf{g}_k e igualando-a a zero, temos, para ($k = 1, \dots, K$):

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{g}_k} &= \sum_{i=1}^n (u_{ik})^m \frac{\partial \mathcal{K}^{(m)}(\mathbf{x}_i, \mathbf{g}_k)}{\partial \mathbf{g}_k} = 0 \\
\sum_{i=1}^n (u_{ik})^m \frac{\partial \exp \left\{ -\frac{(\mathbf{x}_i - \mathbf{g}_k)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{g}_k)}{2\sigma^2} \right\}}{\partial \mathbf{g}_k} &= 0 \\
\left(-\frac{1}{2\sigma^2} \right) \sum_{i=1}^n (u_{ik})^m \exp \left\{ -\frac{(\mathbf{x}_i - \mathbf{g}_k)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{g}_k)}{2\sigma^2} \right\} (-2\mathbf{M}(\mathbf{x}_i - \mathbf{g}_k)) &= 0
\end{aligned}$$

$$\begin{aligned}
\left(\frac{1}{\sigma^2} \right) \sum_{i=1}^n (u_{ik})^m \mathcal{K}^{(m)}(\mathbf{x}_i, \mathbf{g}_k) \mathbf{M}(\mathbf{x}_i - \mathbf{g}_k) &= 0 \\
\left(\frac{1}{\sigma^2} \right) \mathbf{M} \sum_{i=1}^n (u_{ik})^m \mathcal{K}^{(m)}(\mathbf{x}_i, \mathbf{g}_k) \mathbf{x}_i &= \left(\frac{1}{\sigma^2} \right) \mathbf{M} \mathbf{g}_k \sum_{i=1}^n (u_{ik})^m \mathcal{K}^{(m)}(\mathbf{x}_i, \mathbf{g}_k)
\end{aligned}$$

Multiplicando ambos os lados por \mathbf{M}^{-1} obtemos:

$$\mathbf{g}_k = \frac{\sum_{i=1}^n (u_{ik})^m \mathcal{K}^{(m)}(\mathbf{x}_i, \mathbf{g}_k) \mathbf{x}_i}{\sum_{i=1}^n (u_{ik})^m \mathcal{K}^{(m)}(\mathbf{x}_i, \mathbf{g}_k)} \quad \square$$

Na segunda etapa do algoritmo calcularemos as matrizes de covariâncias \mathbf{M} correspondente a cada distância adaptativa proposta nesta dissertação. Para isso devemos encontrar as expressões que minimizam os seguintes critérios de agrupamentos das Eq. (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 $\det(\mathbf{M}) = 1$, cujas matrizes são atualizadas de acordo com as seguintes expressões:*

(a) A matriz de covariância fuzzy completa única que minimiza o critério \mathbf{J} dado pela Eq. (3.8) é atualizada pela expressão:

$$\mathcal{M} = [\det(\mathbf{Q})]^{1/p} (\mathbf{Q})^{-1},$$

$$\mathbf{Q} = \sum_{k=1}^c \mathbf{C}_k \text{ e } \mathbf{C}_k = \sum_{i=1}^n (u_{ik})^m \mathcal{K}^{(m)}(\mathbf{x}_i, \mathbf{g}_k) (\mathbf{x}_i - \mathbf{g}_k) (\mathbf{x}_i - \mathbf{g}_k)^\top$$

(b) A matriz diagonal de covariância fuzzy única que minimiza o critério \mathbf{J} dado pela Eq. (3.9) é atualizada pela expressão:

$$\mathbf{M} = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_p \end{pmatrix} = [\det(\mathbf{Q})]^{1/p} \mathbf{Q}^{-1}, \quad \mathbf{Q} = \sum_{k=1}^c \mathbf{C}_k,$$

Onde $\mathbf{C}_k = \text{diag} \left(\sum_{i=1}^n (u_{ik})^m \mathcal{K}^{(m)}(\mathbf{x}_i, \mathbf{g}_k) (\mathbf{x}_i - \mathbf{g}_k) (\mathbf{x}_i - \mathbf{g}_k)^\top \right)$

(c) A matriz de covariância fuzzy completa diferente para cada grupo que minimiza o critério \mathbf{J} dado pela Eq. (3.10) é atualizada pela expressão:

$$\mathcal{M}_k = [\det(\mathbf{Q}_k)]^{1/p} (\mathbf{Q}_k)^{-1},$$

$$\mathbf{Q}_k = \sum_{k=1}^c \mathbf{C}_k \text{ e } \mathbf{C}_k = \sum_{i=1}^n (u_{ik})^m \mathcal{K}^{(m)}(\mathbf{x}_i, \mathbf{g}_k) (\mathbf{x}_i - \mathbf{g}_k) (\mathbf{x}_i - \mathbf{g}_k)^\top$$

(d) A matriz diagonal de covariância fuzzy diferente para cada grupo que minimiza o critério \mathbf{J} dado pela Eq. (3.11) é atualizada pela expressão:

$$\mathbf{M}_k = \begin{pmatrix} \lambda_{1k} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_{pk} \end{pmatrix} = [\det(\mathbf{Q}_k)]^{1/p} \mathbf{Q}_k^{-1}, \quad \mathbf{Q}_k = \sum_{k=1}^c \mathbf{C}_k,$$

Onde $\mathbf{C}_k = \text{diag} \left(\sum_{i=1}^n (u_{ik})^m \mathcal{K}^{(m)}(\mathbf{x}_i, \mathbf{g}_k) (\mathbf{x}_i - \mathbf{g}_k) (\mathbf{x}_i - \mathbf{g}_k)^\top \right)$

Demonstração: (a) Como o grau de pertinência u_{ik} de cada padrão c em cada grupo P_i , bem como o parâmetro m e os centróides g_i dos grupos correspondentes P_i ($i = 1, \dots, c$) são fixos, podemos reescrever o critério $J1$ como um função de M :

$$2 \sum_{k=1}^c \sum_{i=1}^n (u_{ik})^m \left\{ 1 - \exp \left\{ - \frac{(\mathbf{x}_i - \mathbf{g}_k)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{g}_k)}{2\sigma^2} \right\} \right\}.$$

Seja $g(\mathbf{M}) = 1 - \det(\mathbf{M})$. Queremos determinar os extremos de $J1(\mathbf{M})$ com a restrição $g(\mathbf{M}) = 0$. Seja a função Lagrangiana:

$$\begin{aligned}
L(\mathbf{M}, \beta) &= J1(\mathbf{M}) + \beta g(\mathbf{M}) \\
&= 2 \sum_{k=1}^c \sum_{i=1}^n (u_{ik})^m \left\{ 1 - \exp \left\{ -\frac{(\mathbf{x}_i - \mathbf{g}_k)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{g}_k)}{2\sigma^2} \right\} \right\} + \beta(1 - \det(\mathbf{M})).
\end{aligned}$$

Tomando a sua derivada, temos

$$\frac{\partial L(\mathbf{M}, \beta)}{\partial \mathbf{M}} = \frac{1}{\sigma^2} \sum_{k=1}^c \sum_{i=1}^n (u_{ik})^m \mathcal{K}^{(m)}(\mathbf{x}_i, \mathbf{g}_k) (\mathbf{x}_i - \mathbf{g}_k) (\mathbf{x}_i - \mathbf{g}_k)^\top - \beta \det(\mathbf{M}) \mathbf{M}^{-1}.$$

Nisso, temos que:

$$\beta \det(\mathbf{M}) \mathbf{M}^{-1} = \frac{1}{\sigma^2} \sum_{k=1}^c \sum_{i=1}^n (u_{ik})^m \mathcal{K}^{(m)}(\mathbf{x}_i, \mathbf{g}_k) (\mathbf{x}_i - \mathbf{g}_k) (\mathbf{x}_i - \mathbf{g}_k)^\top.$$

Mas, como $\det(\mathbf{M}) = 1$, temos que

$$\mathbf{M}^{-1} = \frac{1}{\beta \sigma^2} \sum_{k=1}^c \sum_{i=1}^n (u_{ik})^m \mathcal{K}^{(m)}(\mathbf{x}_i, \mathbf{g}_k) (\mathbf{x}_i - \mathbf{g}_k) (\mathbf{x}_i - \mathbf{g}_k)^\top$$

$$\mathbf{M}^{-1} = \frac{1}{\beta \sigma^2} \sum_{k=1}^c \mathbf{C}_k, \quad \mathbf{C}_k = \sum_{i=1}^n (u_{ik})^m \mathcal{K}^{(m)}(\mathbf{x}_i, \mathbf{g}_k) (\mathbf{x}_i - \mathbf{g}_k) (\mathbf{x}_i - \mathbf{g}_k)^\top$$

$$\mathbf{M}^{-1} = \frac{1}{\beta \sigma^2} \mathbf{Q}, \quad \mathbf{Q} = \sum_{k=1}^c \mathbf{C}_k$$

Como $\det(\mathbf{M}^{-1}) = 1/\det(\mathbf{M}) = 1$, a partir de $\mathbf{M}^{-1} = (1/\beta)\mathbf{Q}$ segue que $\det(\mathbf{M}^{-1}) = (1/\beta^p) \det(\mathbf{Q}) = 1 \Rightarrow \beta = (\det(\mathbf{Q}))^{1/p}$. Além disso, como $\mathbf{M}^{-1} = (1/\beta)\mathbf{Q} = (1/(\det(\mathbf{Q}))^{1/p})\mathbf{Q}$, de onde concluimos que $\mathbf{M} = (\det(\mathbf{Q}))^{1/p} \mathbf{Q}^{-1}$. \square

Para os casos (b), (c) e (d) seguimos um raciocínio similar a demonstração utilizada na parte (a) para encontrarmos a matriz \mathbf{M}_k . Podemos verificar que todos são casos especiais do método $J1$, na parte (c) encontramos matrizes de covariâncias completas e diferentes para cada grupo. E por fim as partes (b) e (d) são matrizes diagonais de covariância fuzzy, única e diferente para cada grupo, respectivamente.

Na terceira e última etapa do algoritmo, precisamos agora determinar os graus de pertinências fuzzy u_{ik} ($k = 1, \dots, c; i = 1, \dots, n$). Para isso, os centróides dos grupos $\mathbf{g}_k, k = 1, \dots, c$ são mantidos fixos. O problema se torna, então, encontrar os graus que minimizam o critério J sob restrições adequadas.

Proposição 3.2.3. *O grau de pertinência fuzzy $u_{ik} = (u_{1k}, \dots, u_{ck})$ que minimizam a função objetivo J dada pela Eq. (3.7), sob $u_{ik} \in [0, 1] \forall i$ e $h(u) = \sum_{k=1}^c u_{ik} = 1 \forall i$, são atualizados de acordo com a seguinte equação:*

$$u_{ik} = \left[\sum_{k=1}^c \left(\frac{\varphi^2(\mathbf{x}_i, \mathbf{g}_k)}{\varphi^2(\mathbf{x}_i, \mathbf{g}_k)} \right)^{\frac{1}{m-1}} \right]^{-1} \quad (3.13)$$

Demonstração: Queremos minimizar J com respeito a $u_{ik} = (u_{1k}, \dots, u_{ck})$ sob a restrição de que $u_{ik} \in [0, 1] \forall i$ e $\sum_{k=1}^c u_{ik} = 1 \forall i$, e com os centróides g_k ($k = 1, \dots, c$) e o parâmetro m é mantido fixo. Para isso, utilizaremos o método dos multiplicadores de Lagrange para resolver o seguinte sistema

$$\nabla J(u_{1k}, \dots, u_{ck}) = \mu \nabla h(u_{1k}, \dots, u_{ck})$$

Então, para $i = 1, \dots, c$, temos que :

$$\frac{\partial J(u_{1k}, \dots, u_{ck})}{\partial u_{ik}} = m(u_{ik})^{m-1} \varphi^2(\mathbf{x}_i, \mathbf{g}_k)$$

$$\frac{\partial h(u_{1k}, \dots, u_{ck})}{\partial u_{ik}} = 1$$

$$\frac{\partial J(u_{1k}, \dots, u_{ck})}{\partial u_{ik}} = \mu \frac{\partial h(u_{1k}, \dots, u_{ck})}{\partial u_{ik}}$$

$$m(u_{ik})^{m-1} \varphi^2(\mathbf{x}_i, \mathbf{g}_k) = \mu$$

$$(u_{ik})^{m-1} = \left(\frac{\mu}{m} \right) \cdot \frac{1}{\varphi^2(\mathbf{x}_i, \mathbf{g}_k)}$$

$$u_{ik} = \left(\frac{\mu}{m} \right)^{\frac{1}{m-1}} \cdot \frac{1}{(\varphi^2(\mathbf{x}_i, \mathbf{g}_k))^{\frac{1}{m-1}}} \quad (3.14)$$

Como sabemos que $\mu = 1 \sum_{k=1}^c u_{ik} = 1$, daí temos

$$\begin{aligned} \left(\frac{\mu}{m}\right)^{\frac{1}{m-1}} \cdot \sum_{k=1}^c \frac{1}{(\varphi^2(\mathbf{x}_i, \mathbf{g}_k))^{\frac{1}{m-1}}} &= 1 \\ \left(\frac{\mu}{m}\right)^{\frac{1}{m-1}} &= \left[\sum_{k=1}^c \frac{1}{(\varphi^2(\mathbf{x}_i, \mathbf{g}_k))^{\frac{1}{m-1}}} \right] \end{aligned} \quad (3.15)$$

Resolvendo a equação (3.15) para $\left(\frac{\mu}{m}\right)^{\frac{1}{m-1}}$ e substituindo em (3.14), temos que um extremo de \mathbf{J} é alcançado quando

$$\begin{aligned} u_{ik} &= \frac{1}{\frac{1}{(\varphi^2(\mathbf{x}_i, \mathbf{g}_k))^{\frac{1}{m-1}}}} \cdot \frac{1}{(\varphi^2(\mathbf{x}_i, \mathbf{g}_k))^{\frac{1}{m-1}}} \\ u_{ik} &= \left[\sum_{k=1}^c \left(\frac{\varphi^2(\mathbf{x}_i, \mathbf{g}_k)}{\varphi^2(\mathbf{x}_i, \mathbf{g}_k)} \right)^{\frac{1}{m-1}} \right]^{-1}. \end{aligned}$$

□

Os algoritmos dos métodos de agrupamento fuzzy baseado no Kernel de Mahalanobis para matrizes de covariâncias globais e/ou locais completas e/ou diagonais são 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 (um número máximo de iterações); e fixe $0 < \varepsilon \ll 1$; inicialize os graus de pertinência difusos u_{ik} ($k = 1, \dots, c, i = 1, \dots, n$) de tal forma que $u_{ik} \geq 0 \forall i, k$ e $\sum_{k=1}^c u_{ik} = 1 \forall i$, faça $t = 1$.

2. Etapa 1: Definição dos melhores centróides dos grupos. Atualize os centróides dos grupos \mathbf{g}_k ($k = 1, \dots, c$) de acordo com a Eq (2.6).

3. Etapa 2: Definição das matrizes de covariâncias. Atualize o cálculo das matrizes \mathbf{M} (globais) e \mathbf{M}_k (locais) de acordo com as Eq's. 2.6, ??, ?? e ??,

4. Etapa 3: Definição da melhor partição difusa Atualize os graus de pertinências difusos u_{ik} ($k = 1, \dots, c, i = 1, \dots, n$) de acordo com a Eq. (2.7)

5. Critério de parada Se $|J^{t+1} - J^t| \leq \varepsilon$ ou $t > T$, então, *PARE*, caso contrário, faça $t = t + 1$ e volte ao passo (2).

Capítulo 4

Esquema Experimental

4.1 Introdução

Neste capítulo apresentaremos um esquema de avaliação experimental dos métodos de agrupamento fuzzy baseados no Kernel de Mahalanobis com distâncias quadráticas adaptativas introduzidos no Capítulo 3. Os métodos propostos serão comparados a métodos clássicos de agrupamento como o fuzzy k-médias e sua versão baseada no Kernel Gaussiano, através de experimentos de Monte Carlo considerando quatro configurações de conjuntos de dados simulados. Foram realizadas aplicações com diversos conjuntos de dados reais obtidos do Repositório de Aprendizagem de Máquina da Universidade da Califórnia, Irvine, Estados Unidos (BACHE e LICHMAN [32]).

4.2 Índices de Avaliação

Para comparar os métodos de agrupamento considerados nesta dissertação, utilizamos o Índice Corrigido de Rand (CR) (HUBERT e ARABIE [33]) e a Taxa Total de Erro de Classificação (OERC) (BREIMAN *et al.* [34]).

Seja $P = \{P_1, \dots, P_i, \dots, P_c\}$ 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. As quantidades $n_{ik}, i = 1, \dots, c, k = 1, \dots, K$, representam o número de observações que estão na classe P_i e no grupo P_k e podem ser representados na forma da Tabela 4.1, denominada matriz de confusão.

Tabela 4.1: Matriz de confusão

Classes	Grupos					
	P_1	...	P_k	...	P_K	Σ
P_1	n_{11}	...	n_{1k}	...	n_{1K}	$n_{1.} = \sum_{k=1}^K n_{1k}$
\vdots	\vdots	...	\vdots	...	\vdots	\vdots
P_i	n_{i1}	...	n_{ik}	...	n_{iK}	$n_{i.} = \sum_{k=1}^K n_{ik}$
\vdots	\vdots	...	\vdots	...	\vdots	\vdots
P_c	n_{c1}	...	n_{ck}	...	n_{cK}	$n_{c.} = \sum_{k=1}^K n_{ck}$
Σ	$n_{.1} = \sum_{i=1}^c n_{i1}$...	$n_{.k} = \sum_{i=1}^c n_{ik}$...	$n_{.K} = \sum_{i=1}^c n_{iK}$	$n = \sum_{i=1}^c \sum_{k=1}^K n_{ik}$

O Índice Corrigido de Rand é obtido como

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

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

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 (ARABIE *et al.* [35]).

Em problemas de classificação, cada grupo P_k é associado a uma classe *a priori* P_i e esta associação deve ser interpretada como se a verdadeira classe *a priori* fosse 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 P_i . Para obter uma taxa de erro de classificação mínima, precisamos encontrar uma regra de decisão que minimize a probabilidade de erro.

Seja $\ell(P_i, P_k)$ a probabilidade *a posteriori* de que uma observação pertença à classe P_i quando associado 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* $\ell(P_i, P_k)$ e o índice da classe *a priori* associada a esta moda é dada por:

$$MAP(P_k) = \arg \max_{1 \leq i \leq c} \ell(P_i, P_k)$$

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 alocação do grupo P_k é igual a $1 - \ell(P_{MAP(P_k)}/P_k)$ e a taxa total de erro de classificação (OERC) é igual a

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

Para uma amostra,

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

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:

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

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*.

4.3 Conjunto de Dados Simulados

Para cada conjunto de dados sintéticos simulamos classes com tamanhos e formas diferentes. Os conjuntos de dados simulados terão 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 covariância representada por

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

Nisso iremos considerar quatro configurações de conjuntos de dados simulados:

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

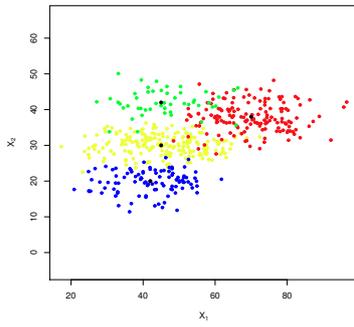
2. As matrizes de covariância dos grupos são diagonais e diferentes para cada grupo;
3. As matrizes de covariância dos grupos são completas e aproximadamente iguais;
4. As matrizes de covariância dos grupos são completas e diferentes para cada grupo.

A Tabela 4.2 apresenta as configurações dos dados simulados para cada classe e seus respectivos parâmetros, bem como os valores dos vetores de média e as matrizes de covariância (formadas pelas variâncias e as correlações de cada classe) para cada conjuntos de dados simulados.

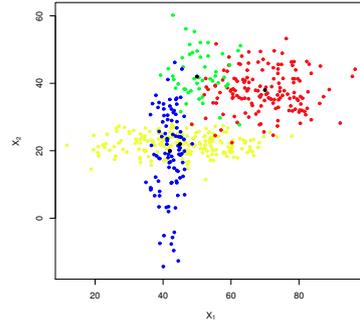
Tabela 4.2: Configurações dos Cenários

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

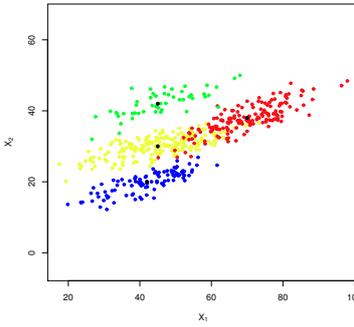
A Figura 4.1 ilustra as configurações dos quatros conjuntos de dados simulados em que consideramos: (a) As matrizes de covariância diagonais e globais (quase iguais); (b) As matrizes de covariância diagonais e locais (diferente para cada grupo); (c) As matrizes de covariância não diagonais (completas) e globais; (d) As matrizes de covariância não diagonais e locais.



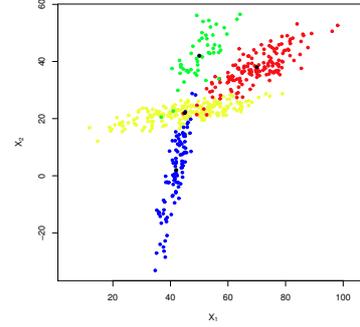
(a) Cenário 1



(b) Cenário 2



(c) Cenário 3



(d) Cenário 4

Figura 4.1: Conjuntos de Dados Simulados.

4.4 Avaliação com Dados Simulados

O desempenho dos métodos de agrupamento fuzzy será avaliado de acordo com os índices CR e OERC calculados a partir da partição fuzzy obtida de uma simulação de Monte Carlo com 100 repetições para cada configuração de dados. Cada replicação fornece uma matriz de confusão, em que os grupos fornecidos pelo algoritmo de agrupamento são colocados em linhas e as classes *a priori* são colocadas em colunas.

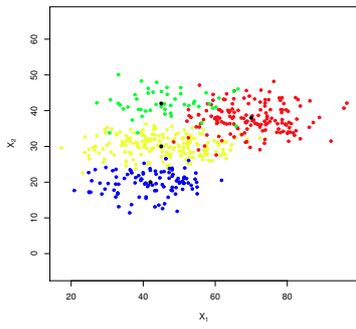
Em cada replicação, os algoritmos de agrupamento foram executados 100 vezes até convergir para um valor estacionário do critério de adequação, e depois disso o melhor resultado para cada algoritmo foi selecionado de acordo com o critério de adequação. A média e o desvio padrão destas medidas foram computados. Os algoritmos foram aplicados aos conjuntos de dados simulados 1, 2, 3 e 4 para obtenção de 4 grupos. As partições dos 4 grupos fornecidas pelos algoritmos de agrupamento foram comparadas a partição *a priori* conhecidas de 4 classes.

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.

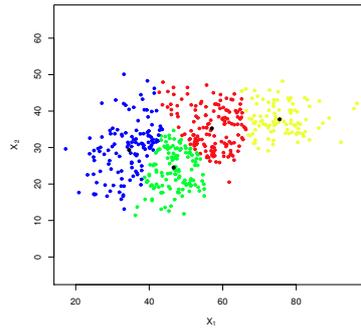
Algoritmos	CR	OERC
FCM	0.303 (0.050)	0.418 (0.044)
KFCM	0.321 (0.060)	0.405 (0.049)
MK-FCM-GD	0.623 (0.105)	0.197 (0.081)
MK-FCM-GF	0.617 (0.095)	0.198 (0.072)
MK-FCM-LD	0.588 (0.096)	0.227 (0.078)
MK-FCM-LF	0.554 (0.096)	0.246 (0.075)

A Tabela 4.3 apresenta o desempenho dos algoritmos FCM, KFCM, MK-FCM-GD, MK-FCM-GF, MK-FCM-LD e MK-FCM-LF em valores de média e desvio padrão para o conjunto de dados simulados 1, de acordo com os índices CR e OERC. Verifica-se que os algoritmos globais baseados em uma única distância adaptativa (MK-FCM-GD e MK-FCM-GF) superam os algoritmos com base em distâncias adaptativas locais (MK-FCM-LD e MK-FCM-LF) e os métodos clássicos (FCM e KFCM). Analisando esses dois algoritmos que obtiveram bom desempenho e resultados próximos, o algoritmo de matriz de covariância diagonal global (MK-FCM-GD) demonstrou melhor explanação do conjunto de dados 1 com o maior Índice Corrigido de Rand (0.623) e o menor Erro de Taxa Total de Classificação (0.197). E o algoritmo FCM apresentou o pior desempenho com o menor ICR (0.303) e o maior OERC (0.418).

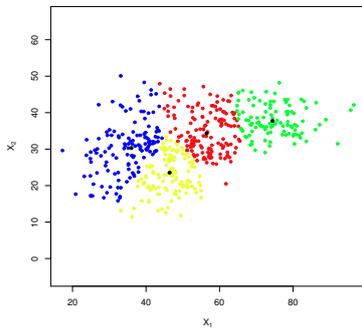
A Figura 4.2 ilustra a configuração do conjunto de dados simulados 1 e mostram os resultados dos agrupamentos fuzzy fornecidos, respectivamente, pelos algoritmos FCM, KFCM, MK-FCM-GD, MK-FCM-GF, MK-FCM-LD e MK-FCM-LF. Como pode ser visto nas Figuras 4.2(d) e 4.2(e), os algoritmos MK-FCM-GD e MK-FCM-GF apresentaram as melhores categorização do conjuntos de dados simulados 1, destacando-se a Figura 4.2(d) cujo algoritmo utiliza uma única matriz de covariância diagonal global que representa a melhor classificação entre os dados simulados do conjunto 1.



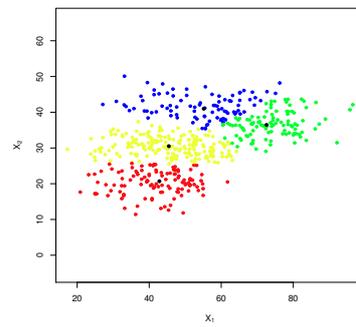
(a) Configuração 1



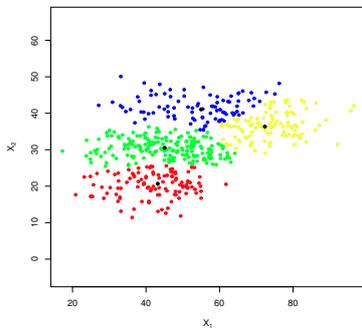
(b) FCMMeans



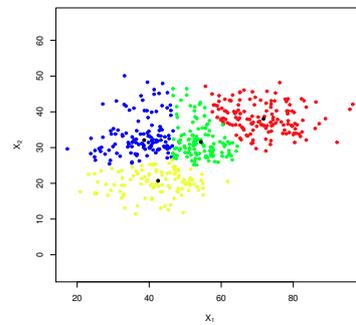
(c) KFCMeans



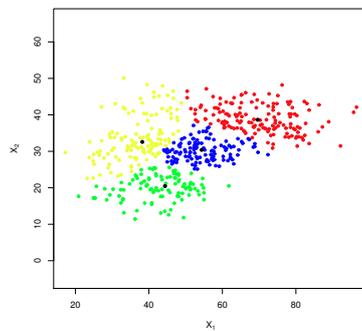
(d) MK-FCM-GD



(e) MK-FCM-GF



(f) MK-FCM-LD



(g) MK-FCM-LF

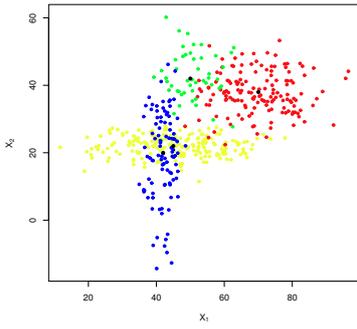
Figura 4.2: Conjuntos de Dados Simulados do 1º Cenário.

O desempenho dos métodos de agrupamentos fuzzy para o conjunto de dados simulados 2 é apresentado na Tabela 4.4. Observar-se que os algoritmos baseados em distâncias adaptativas locais, ou seja, que possuem matrizes de covariâncias distintas para cada grupo (MK-FCM-LD e MK-FCM-LF) fornecem melhores resultados do que os algoritmos baseados em distâncias adaptativas globais (MK-FCM-GD e MK-FCM-GF) e dos métodos clássicos (FCM e KFCM). Considerando esses dois algoritmos que obtiveram bom desempenho, o algoritmo de matriz de covariância diagonal local (MK-FCM-LD) demonstrou melhor explicação do conjunto de dados 2 com o maior Índice Corrigido de Rand (0.535) e o menor Erro de Taxa Total de Classificação (0.244). O algoritmo FCM apresentou a pior performance novamente nesta configuração, com o menor IRC(0.397) e o maior OERC(0.345) entre os algoritmos.

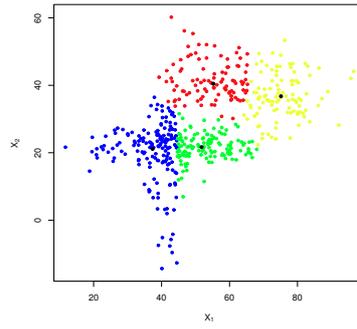
Tabela 4.4: Desempenho dos algoritmos de agrupamento fuzzy no conjunto de dados simulados 2: média e desvio-padrão (entre parênteses) dos índices CR e OERC.

Algoritmos	CR	OERC
FCM	0.397 (0.034)	0.345 (0.030)
KFCM	0.398 (0.038)	0.339 (0.037)
MK-FCM-GD	0.415 (0.033)	0.329 (0.029)
MK-FCM-GF	0.401 (0.032)	0.314 (0.036)
MK-FCM-LD	0.535 (0.062)	0.244 (0.049)
MK-FCM-LF	0.467 (0.097)	0.293 (0.072)

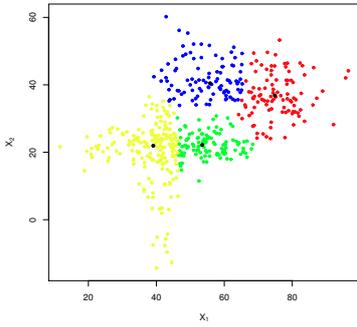
A Figura 4.3 mostra o desempenho dos algoritmos de agrupamento fuzzy no conjunto de dados simulados 2. As Figuras 4.3(b), 4.3(c), 4.3(d), 4.3(e), 4.3(f) e 4.3(g) mostram os resultados de agrupamentos fornecidos, respectivamente pelos algoritmos FCM, KFCM, MK-FCM-GD, MK-FCM-GF, MK-FCM-LD e MK-FCM-LF e a configuração do conjunto de dados simulados 2 pode ser visto na Figura 4.3(a). Observa-se uma melhor categorização dos dados no algoritmo MK-FCM-LD (Figura 4.3(f)) que fornece as matrizes de covariância diagonais locais.



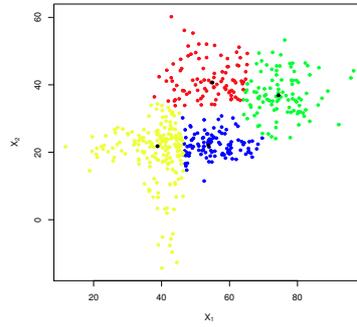
(a) Configuração 2



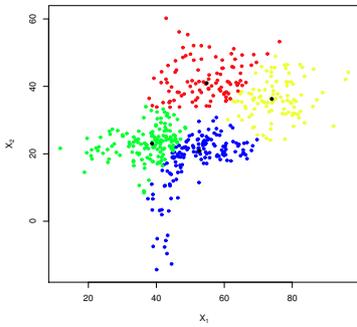
(b) FCM



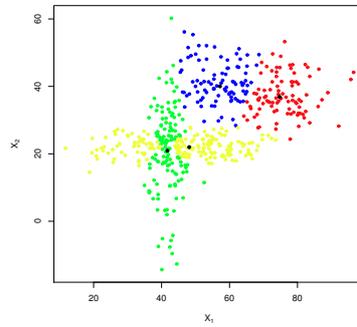
(c) KFCM



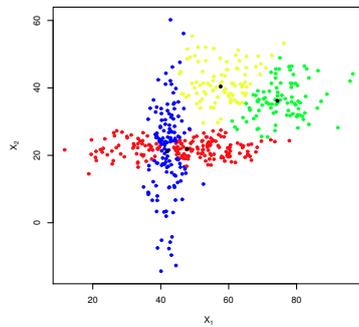
(d) MK-FCM-GD



(e) MK-FCM-GF



(f) MK-FCM-LD



(g) MK-FCM-LF

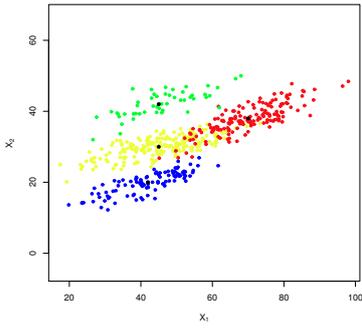
Figura 4.3: Conjuntos de Dados Simulados do 2º Cenário.

O desempenho dos métodos de agrupamento fuzzy para o conjunto de dados simulados 3 é apresentado na Tabela 4.5. Percebe-se que dois algoritmos se destacaram claramente dos demais por apresentarem resultados superiores, são eles os de características não diagonais: global (MK-FCM-GF) e local (MK-FCM-LF), ou seja, apresentam as matrizes de covariâncias completas. Entre os algoritmos que explicaram melhor o conjuntos de dados 3, o algoritmo MK-FCM-GF se destacou possuindo o maior Índice Corrigido de Rand (0.758) e o menor Erro de Taxa Total de Classificação (0.083). O algoritmo FCM apresentou o pior desempenho com o menor ICR (0.252) e o maior OERC (0.446).

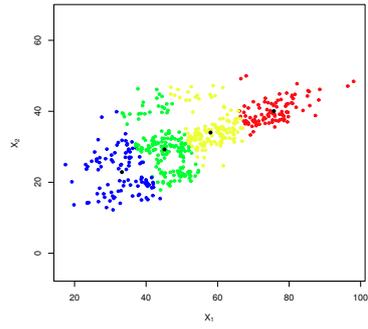
Tabela 4.5: Desempenho dos algoritmos de agrupamento fuzzy no conjunto de dados simulados 3: média e desvio-padrão (entre parênteses) dos índices CR e OERC.

Algoritmos	CR	OERC
FCM	0.252 (0.028)	0.446 (0.033)
KFCM	0.264 (0.042)	0.441 (0.041)
MK-FCM-GD	0.452 (0.029)	0.311 (0.022)
MK-FCM-GF	0.758 (0.031)	0.083 (0.012)
MK-FCM-LD	0.435 (0.052)	0.328 (0.036)
MK-FCM-LF	0.756 (0.029)	0.084 (0.011)

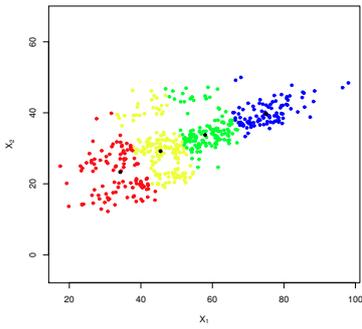
A Figura 4.4 ilustra a configuração do conjunto de dados simulados 3 e mostram os resultados dos agrupamentos fuzzy fornecidos, respectivamente, pelos algoritmos FCM, KFCM, MK-FCM-GD, MK-FCM-GF, MK-FCM-LD e MK-FCM-LF. Como pode ser visto nas Figuras 4.4(e) e 4.4(g), os algoritmos MK-FCM-GF e MK-FCM-LF apresentaram as melhores categorização do conjuntos de dados simulados 3, destacando-se a Figura 4.4(e) cujo algoritmo utiliza uma matriz de covariância não diagonal que representa a melhor classificação entre os dados simulados do conjunto 3. É possível verificar que os grupos estão dispostos com uma leve inclinação, o que indica direção, comprovando assim a existência de covariância entre as variáveis.



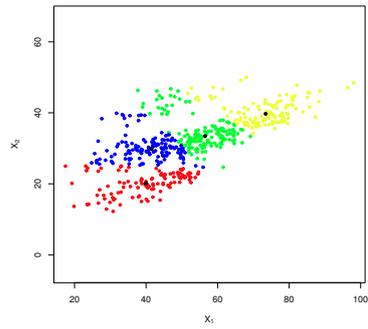
(a) Configuração 3



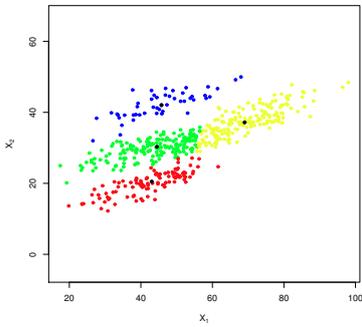
(b) FCM



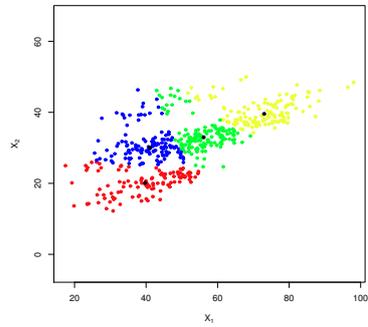
(c) KFCM



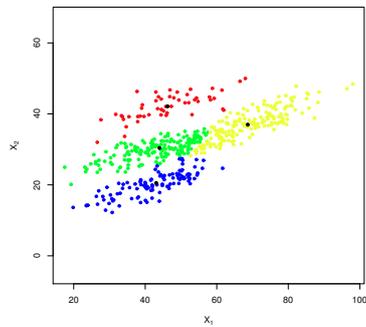
(d) MK-FCM-GD



(e) MK-FCM-GF



(f) MK-FCM-LD



(g) MK-FCM-LF

Figura 4.4: Conjuntos de Dados Simulados do 3º Cenário.

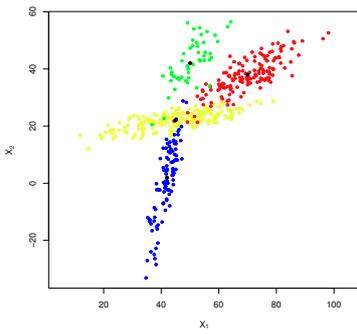
O desempenho dos métodos de agrupamento fuzzy para o conjunto de dados simulados 4, de acordo com os índices CR e OERC, é apresentado na Tabela 4.6. Verifica-se que o único algoritmo que mostrou melhor desempenho entre os demais nesse cenário foi o baseado em distância adaptativa local e não diagonal (MK-FCM-LF), com o maior Índice Corrigido de Rand (0.783) e o menor Erro de Taxa Total de Classificação (0.099). O algoritmo FCM apresentou a pior performance novamente nesta configuração, com o menor IRC(0.416) e o maior OERC(0.329) entre os algoritmos.

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.

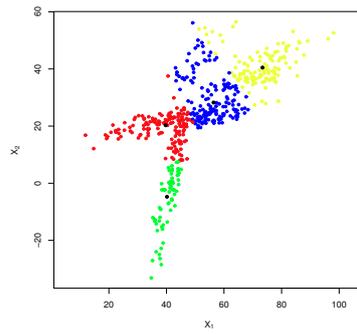
Algoritmos	CR	OERC
FCM	0.416 (0.028)	0.329 (0.025)
KFCM	0.430 (0.025)	0.322 (0.022)
MK-FCM-GD	0.433 (0.024)	0.325 (0.022)
MK-FCM-GF	0.422 (0.026)	0.328 (0.019)
MK-FCM-LD	0.527 (0.037)	0.267 (0.028)
MK-FCM-LF	0.783 (0.124)	0.099 (0.078)

A Figura 4.5 mostra o desempenho dos algoritmos de agrupamento fuzzy no conjunto de dados simulados 4. As Figuras 4.5(b), 4.5(c), 4.5(d), 4.5(e), 4.5(f) e 4.5(g) mostram os resultados de agrupamentos fornecidos, respectivamente pelos algoritmos FCM, KFCM, MK-FCM-GD, MK-FCM-GF, MK-FCM-LD e MK-FCM-LF e a configuração do conjunto de dados simulados 4 pode ser visto na Figura 4.5(a). Observa-se uma melhor categorização dos dados no algoritmo MK-FCM-LF (Figura 4.5(g)) que fornece as matrizes de covariância locais não diagonais. Assim como nas Figuras 4.4(b), 4.4(c), 4.4(d), 4.4(e), 4.4(f) e 4.4(g), os grupos estão levemente inclinados, característica da relação de covariância entre as variáveis.

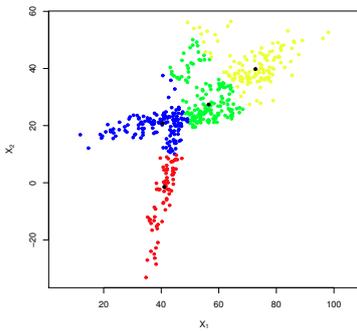
Os algoritmos propostos foram aplicados a quatro conjunto de dados simulados, e em relação a esses dados simulados o desempenho dos algoritmos de agrupamentos fuzzy baseados em distâncias adaptativas (MK-FCM-GD, MK-FCM-GF, MK-FCM-LD e MK-FCM-LF) superam os métodos clássicos FCM e KFCM, tanto quando as matrizes de covariâncias são diagonais globais (matriz única) e locais (matriz diferente para cada grupo) quanto quando são não diagonais globais e locais. Em todos os conjuntos o Kernel de Mahalanobis foi superior em relação a contrapartida.



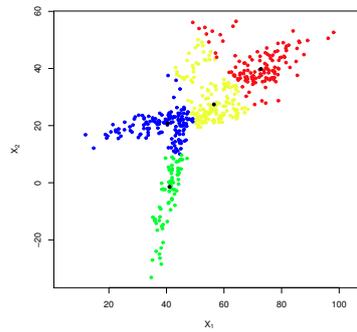
(a) Configuração 4



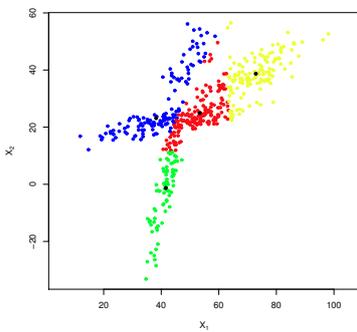
(b) FCM



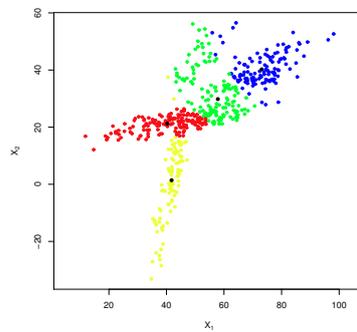
(c) KFCM



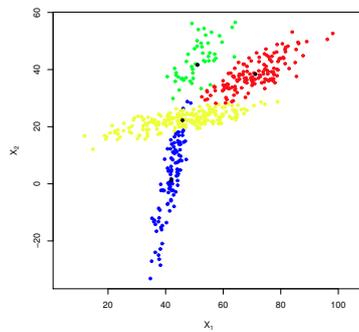
(d) MK-FCM-GD



(e) MK-FCM-GF



(f) MK-FCM-LD



(g) MK-FCM-LF

Figura 4.5: Conjuntos de Dados Simulados do 4º Cenário.

4.5 Conjuntos de Dados Reais

Foram também 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*. A Tabela 4.7 (na qual n representa o número de padrões, p representa o número de variáveis e c representa o número de classes a priori) apresenta as dimensões dos conjuntos de dados reais considerados. Nesta seção também descreveremos algumas informações a cerca dos conjuntos da Tabela 4.7, que podem ser visto em BACHE e LICHMAN [32].

Tabela 4.7: 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

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.

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 seqüência de sinal clivável, outras proteínas da membrana externa, proteínas periplasmáticas, brãia interna proteínas com uma seqüência de sinal não-clivável, proteínas da membrana interna sem uma seqüência 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 padrões obtidos por sinais de sonar captados em um cilindro de metal em vários ângulos e sob várias condições. O sonar das rochas contém 97 padrões 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 em 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.

4.6 Avaliação com Dados Reais

Os algoritmos de agrupamento FCM, KFCM, MK-FCM-GD, MK-FCM-GF, MK-FCM-LD e MK-FCM-LF foram aplicados em conjuntos de dados reais da Tabela 4.7. 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 centróides) e selecionou-se o melhor resultado de acordo com os critério de adequação.

Os critérios de adequação utilizados foram o Índice Corrigido de Rand (ICR) e o Erro de Taxa Total de Classificação (OERC), que foram calculados para as melhores soluções. O parâmetro m foi configurado igual a 2. 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 4.8. Para cada conjunto será destacado os algoritmos que obtiveram melhores desempenho perante os demais (estão descritos em negrito), ou seja, os algoritmos que apresentarem maior índice CR e menor OERC.

Nisso observamos que os algoritmos baseados no Kernel de Mahalanobis proposto nesta dissertação apresentaram desempenho superior aos métodos clássicos nos seguintes conjuntos de dados reais: *Breast Tissue*, *Íris*, *Thyroid Disease*, *WDBC* e *Wine*. Nos conjuntos *Ecoli* e *Rukiln* os algoritmos baseados em matrizes de covariâncias locais (MK-FCM-LF e MK-FCM-LD) apresentaram os melhores desempenhos juntamente com o método clássico FCM. Nos conjuntos *Vertebral Column* e *Sonar* os algoritmos clássicos foram os que apresentaram bons desempenhos juntamente com os algoritmos de matrizes de covariância não diagonais (MK-FCM-GF e MK-FCM-LF).

Os métodos de agrupamentos fuzzy com matrizes de covariâncias diagonais (MK-FCM-GD e MK-FCM-LD) se adequaram bem aos conjuntos de dados *Breast Tissue*, *WDBC* e *Wine*. Destacando dentre os novos o conjunto de dados *Íris* se adequou bem aos algoritmos de matrizes de covariâncias globais (MK-FCM-GD e MK-FCM-GF) e o conjunto de dados *Thyroid Disease* se adequou bem aos algoritmos de matrizes de covariâncias locais (MK-FCM-LD e MK-FCM-LF).

Avaliamos neste capítulo o desempenho dos métodos de agrupamento fuzzy baseados no Kernel de Mahalanobis introduzidos no Capítulo 3 através de experimentos com dados simulados (simulação de Monte Carlos) e dados reais. Os resultados numéricos dos experimentos de Monte Carlos evidenciaram a superioridade dos algoritmos de agrupamentos fuzzy. E os experimentos com dados reais neste capítulo mostraram a utilidade dos algoritmos propostos em diversas situações práticas da realidade.

Tabela 4.8: 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	0.169	0.556	0.326	0.309	0.338	0.524
MK-FCM-GD	0.298	0.471	0.248	0.374	0.348	0.518
MK-FCM-GF	0.258	0.491	0.311	0.316	0.336	0.524
MK-FCM-LD	0.278	0.481	0.261	0.364	0.297	0.538
MK-FCM-LF	0.189	0.518	0.120	0.458	0.387	0.422
<i>Algoritmos</i>	<i>Íris</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	0.757	0.093	0.002	0.407	0.001	0.461
MK-FCM-GD	0.868	0.046	0.137	0.296	-0.004	0.476
MK-FCM-GF	0.867	0.045	0.206	0.259	-0.003	0.475
MK-FCM-LD	0.866	0.044	0.709	0.074	-0.004	0.466
MK-FCM-LF	0.743	0.100	0.284	0.222	-0.005	0.466
<i>Algoritmos</i>	<i>Thyroid Disease</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	0.128	0.409	0.550	0.126	0.375	0.292
MK-FCM-GD	0.152	0.409	0.773	0.059	0.835	0.056
MK-FCM-GF	0.002	0.609	0.074	0.362	0.341	0.354
MK-FCM-LD	0.877	0.037	0.818	0.047	0.742	0.089
MK-FCM-LF	0.601	0.135	0.677	0.088	0.169	0.387

Capítulo 5

Conclusões

Nesta dissertação propusemos métodos de agrupamento fuzzy baseados no Kernel de Mahalanobis com distâncias quadráticas adaptativas, em que comparamos aos métodos de agrupamentos baseados em Kernel clássicos, a abordagem proposta possui a vantagem de possibilitar a utilização das distâncias quadráticas adaptativas, que mudam a cada iteração dos algoritmos e podem tanto ser a mesma para todos os grupos (distâncias adaptativas globais), quanto serem diferentes de um grupo para outro (distâncias adaptativas locais).

Para cada algoritmo, obtivemos uma expressão para os melhores centroídes dos grupos e para os graus de pertinência fuzzy por meio do Kernel de Mahalanobis. Obtivemos também as expressões das matrizes de covariância fuzzy para cada método por meio da extração de sua derivada da função objetivo correspondente para cada matriz de covariância. A utilidade dos métodos propostos foi demonstrada através de experimentos numéricos com conjuntos de dados simulados por meio de simulações de Monte Carlos e conjuntos de dados reais.

Esses conjuntos foram avaliados de acordo com os Índice Corrigido de Rand e a Taxa Total de Erro de Classificação. Com relação aos conjuntos de dados simulados, os métodos de agrupamento fuzzy baseado em distâncias quadráticas adaptativas globais e locais com matrizes de covariância diagonais e/ou não diagonais apresentaram desempenho superior aos métodos clássicos durante as simulações dos quatro cenários de dados. Em todos os conjuntos o Kernel de Mahalanobis foi superior em relação a contrapartida.

Destacamos assim os algoritmos que apresentaram bons desempenhos nas simulações para cada cenário. Durante a simulação do 1º cenário se destacou o método MK-FCM-GD, no 2º cenário foi o método MK-FCM-LD, no 3º cenário o algoritmo que apresentou resultados superiores aos demais foi o MK-FCM-GF e por fim o 4º cenário, o método MK-FCM-LF foi superior aos demais algoritmos. E os piores desempenhos para os 4 conjuntos de dados simulados foram apresentados pelo algoritmo k-médias fuzzy clássico (FCM).

Para o conjunto de dados reais, o desempenho dos algoritmos de agrupamento fuzzy foi avaliado de acordo com os índices CR e OERC. De maneira geral, as melhores classificações de desempenho foram obtidas pelos algoritmos MKFCM-GD e MKFCMLD. Nos conjuntos *Breast Tissue*, *WDBC* e *Wine*, os dois métodos que se destacaram entre os resultados mostrados foram os de matrizes de covariância diagonais (MK-FCM-GD e MK-FCM-LD).

No conjunto *Íris* os dois algoritmos cujos resultados apresentaram desempenho superior aos outros foram os de matrizes de covariâncias globais (MK-FCM-GD e MK-FCM-GF). Ao contrário do conjunto *Íris*, os algoritmos que apresentaram bons resultados para o conjunto *Thyroid Disease* foram os de matrizes de covariâncias locais (MK-FCM-LD e MK-FCM-LF). Nos conjuntos *Vertebral Column*, *Ecoli*, *Rukiln* e *Sonar* os métodos clássicos FCM e K-FCM juntamente com os de matrizes de covariância locais e globais (MK-FCM-LF, MK-FCM-LD e MK-FCM-GF) apresentaram bons desempenhos durante as simulações.

Em virtude dos resultados apresentados, recomendamos a utilização dos métodos de agrupamento fuzzy baseado no Kernel de Mahalanobis através de distâncias adaptativas locais (MK-FCM-LF e MK-FCM-LD). Esses métodos apresentaram, na maioria das situações experimentais consideradas nesta dissertação de mestrado, desempenhos superiores aos desempenhos dos demais métodos de agrupamento.

Referências Bibliográficas

- [1] JAIN, A. K. “Data clustering: 50 years beyond K-means”, *Pattern Recogn. Lett.*, v. 31, n. 8, pp. 651–666, jun. 2010. ISSN: 0167-8655. doi: 10.1016/j.patrec.2009.09.011. Disponível em: <<http://dx.doi.org/10.1016/j.patrec.2009.09.011>>.
- [2] VICINI, L. *Análise multivariada da teoria à prática*. monografia, Universidade Federal de Santa Maria, 2005.
- [3] MANLY, B. F. J. *Métodos Estatísticos Multivariados: uma introdução*. 3th ed. , Bookman, 2008. ISBN: 978-85-7780-185-5.
- [4] MINGOTI, S. A. *Métodos de Agrupamentos para Dados Quantitativos e Qualitativos*. 1th ed. , EST-UFMG, 2013.
- [5] HÖPPNER, F., KLAWONN, F., KRUSE, R., et al. *Fuzzy Cluster Analysis*. John Wiley & Sons, Inc., 1999. ISBN: 3-540-40317-5.
- [6] FERREIRA, M., DE CARVALHO, F. “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 2014. doi: 10.1109/IJCNN.2014.6889653.
- [7] ZHANG, D. Q., CHEN, S. C. “Fuzzy clustering using kernel method”. In: *The 2002 International Conference on Control and Automation, 2002 ICCA*, pp. 162–163, 2002.
- [8] ZHANG, D. Q., CHEN, S. C., PAN, Z. S., et al. “Kernel-based fuzzy clustering incorporating spatial constraints for image segmentation”. In: *International Conference on Machine Learning and Cybernetics*, v. 4, pp. 2189–2192, 2003.
- [9] CHEN, S. C., ZHANG, D. Q. “Robust Image Segmentation Using FCM with Spatial Constraints Based on New Kernel-induced Distance Measure”, *IEEE Trans. Systems Man Cybernet*, v. 34, n. 4, pp. 1907–1916, 2004.

- [10] ZHANG, D. Q., CHEN, S. C. “A novel kernelized fuzzy c -means algorithm with application in medical image segmentation”, *Artif. Intell. Med.*, v. 32, n. 1, pp. 37–50, 2004.
- [11] MACDONALD, D., FYFE, C. “The kernel self-organizing map”. In: *Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies*, v. 1, pp. 317–320, 2000.
- [12] INOKUCHI, R., MIYAMOTO, S. “LVQ clustering and SOM using a kernel function”. In: *Proceedings of IEEE International Conference on Fuzzy Systems*, v. 3, pp. 1497–1500, 2004.
- [13] KIM, D. W., LEE, K. Y., LEE, D., et al. “A kernel-based subtractive clustering method”, *Pattern Recognition Letters*, v. 26, pp. 879–891, 2005.
- [14] QINAND, A. K., SUGANTHAN, P. N. “Kernel neural gas algorithms with application to cluster analysis”. In: *ICPR – 17th International Conference on Pattern Recognition (ICPR’04)*, v. 4, pp. 617–620, 2004.
- [15] FILIPPONE, M., CAMASTRA, F., MASULLI, F., et al. “A survey of kernel and spectral methods for clustering”, *Pattern Recognition*, v. 41, pp. 176–190, 2008.
- [16] CAMASTRA, F., VERRI, A. “A Novel Kernel Method for Clustering”, *IEEE Trans. Neural Networks*, v. 27, n. 5, pp. 801–804, 2005.
- [17] BEN-HUR, A., HORN, D., SIEGELMANN, H. T., et al. “Support Vector Clustering”, *J. Mach. Learn. Res.*, v. 2, pp. 125–137, 2001.
- [18] GRAVES, D., PEDRYCZ, W. “Kernel-based fuzzy clustering and fuzzy clustering: A comparative experimental study”, *Fuzzy Sets and Systems*, v. 161, pp. 522–543, 2010.
- [19] GIROLAMI, M. “Mercer Kernel-based clustering in feature space”, *IEEE Trans. Neural Networks*, v. 13, n. 3, pp. 780–784, 2002.
- [20] BEZDEK, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York, Plenum, 1981.
- [21] KOHONEN, T. “Essentials of the self-organizing map”, *Neural Netw.*, v. 37, pp. 52–65, jan. 2013. ISSN: 0893-6080. doi: 10.1016/j.neunet.2012.09.018. Disponível em: <<http://dx.doi.org/10.1016/j.neunet.2012.09.018>>.

- [22] CAMPS-VALLS, G., RODRIGO-GONZÁLEZ, A., MUÑOZ-MARÍ, J., et al. “Hyperspectral image classification with mahalanobis relevance vector machines.” In: *IGARSS*, pp. 3802–3805, 2007.
- [23] ABE, S. “Training of Support Vector Machines with Mahalanobis Kernels.” In: Duch, W., Kacprzyk, J., Oja, E., et al. (Eds.), *ICANN (2)*, v. 3697, *Lecture Notes in Computer Science*, pp. 571–576. Springer, 2005.
- [24] KAMADA, Y., ABE, S. “Support Vector Regression Using Mahalanobis Kernels”. In: *ANNPR*, pp. 144–152, 2006.
- [25] WANG, D., YEUNG, D. S., TSANG, E. C. C. “Weighted Mahalanobis Distance Kernels for Support Vector Machines”, *IEEE Transactions on Neural Networks*, v. 18, n. 5, pp. 1453–1462, 2007.
- [26] MACQUEEN, J. B. “Some methods of classification and analysis of multivariate observations”. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.
- [27] BEZDEK, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA, Kluwer Academic Publishers, 1981. ISBN: 0306406713.
- [28] MERCER, J. “Functions of Positive and Negative Type and Their connection With the Theory of Integrals Equations”. In: *Proc. R. Soc. London*, v. 209, pp. 415–446, 1909.
- [29] HAYKIN, S. *Neural Networks: A Comprehensive Foundation*. 2nd ed. Upper Saddle River, NJ, USA, Prentice Hall PTR, 1998. ISBN: 0132733501.
- [30] SCHÖLKOPF, B., SMOLA, A. J., MÜLLER, K. R. “Nonlinear component analysis as a kernel eigenvalue problem”, *Neural Comput.*, v. 10, n. 5, pp. 1299–1319, 1998.
- [31] MÜLLER, K. R., MIKA, S., RÄTSCH, G., et al. “An introduction to kernel-based learning algorithms”, *IEEE Trans. Neural Networks*, v. 12, n. 2, pp. 181–202, 2001.
- [32] BACHE, K., LICHMAN, M. “UCI Machine Learning Repository”. 2013. Disponível em: <<http://archive.ics.uci.edu/ml>>.
- [33] HUBERT, L., ARABIE, P. “Comparing partitions”, *Journal of Classification*, v. 2, pp. 193–218, 1985.

- [34] BREIMAN, L., FRIEDMAN, J., STONE, C. J., et al. *Classification and Regression Trees*. Boca Raton, Chapman & Hall/CRC, 1984.
- [35] Arabie, P., Hubert, L., De Soete, G. (Eds.). "Clustering and Classification". cap. Clustering validation: results and implications for applied analysis, pp. 341–375, Singapore, World Scientific, 1996.