

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO



**REPRESENTAÇÃO DO ESPAÇO DE CARACTERÍSTICAS
POR MEIO DE CONJUNTOS DIFUSOS**

ALAN CARLOS GENARI

Uberlândia - Minas Gerais

2010

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO



ALAN CARLOS GENARI

REPRESENTAÇÃO DO ESPAÇO DE CARACTERÍSTICAS POR MEIO DE CONJUNTOS DIFUSOS

Dissertação de Mestrado apresentada à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como parte dos requisitos exigidos para obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Banco de Dados.

Orientadora:

Prof^a. Dr^a. Denise Guliato

Uberlândia, Minas Gerais
2010

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

G324r
2010 Genari, Alan Carlos, 1974-
 Representação do espaço de características por meio de conjuntos
 difusos / Alan Carlos Genari. - 2010.
 65 f. : il.

 Orientadora: Denise Guliato.
 Dissertação (mestrado) -- Universidade Federal de Uberlândia,
Programa de Pós-Graduação em Ciência da Computação.
 Disponível em: <http://dx.doi.org/10.14393/ufu.di.2017.11>
 Inclui bibliografia.

 1. Computação - Teses. 2. Conjuntos difusos - Teses. 3.
Processamento de imagens - Teses. 4. Sistemas de recuperação da
informação - Teses. I. Guliato, Denise. II. Universidade Federal de
Uberlândia. Programa de Pós-Graduação em Ciência da Computação.
III. Título.

CDU: 681.3

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Os abaixo assinados, por meio deste, certificam que leram e recomendam para a Faculdade de Computação a aceitação da dissertação intitulada “**Representação do espaço de características por meio de conjuntos difusos**” por **Alan Carlos Genari** como parte dos requisitos exigidos para a obtenção do título de **Mestre em Ciência da Computação**.

Uberlândia, 17 de Setembro de 2010

Orientadora:

Prof^a. Dr^a. Denise Guliato
Universidade Federal de Uberlândia

Banca Examinadora:

Prof. Dr. Ilmério Reis da Silva
Universidade Federal de Uberlândia

Prof^a. Dr^a. Agma Juci Machado Traina
Universidade de São Paulo

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Data: Setembro de 2010

Autor: **Alan Carlos Genari**
Título: **Representação do espaço de características por meio de conjuntos difusos**
Faculdade: **Faculdade de Computação**
Grau: **Mestrado**

Fica garantido à Universidade Federal de Uberlândia o direito de circulação e impressão de cópias deste documento para propósitos exclusivamente acadêmicos, desde que o autor seja devidamente informado.

Autor

O AUTOR RESERVA PARA SI QUALQUER OUTRO DIREITO DE PUBLICAÇÃO DESTE DOCUMENTO, NÃO PODENDO O MESMO SER IMPRESSO OU REPRODUZIDO, SEJA NA TOTALIDADE OU EM PARTES, SEM A PERMISSÃO ESCRITA DO AUTOR.

Dedicatória

*Aos meus pais Wilson Roberto e Dalva (in memoriam).
À minha esposa Cláudia e minhas filhas Camila e Helena.*

Agradecimentos

Agradeço...

A Deus, por minha vida e por me dar sabedoria, perseverança e calma para enfrentar todos os desafios.

A meus pais Wilson Roberto Genari e Dalva Helena Genari (*in memoriam*) pela minha educação, pela dedicação, apoio, confiança, carinho e amor em todos os momentos.

À minha esposa Cláudia por ter me apoiado tanto durante este projeto. Você me deu muita força, tranquilidade, paz, alegria e amor. E preencheu muitas vezes junto a nossas filhas, a minha ausência. Te amo.

Às minhas filhas Camila e Helena que deram um novo sentido a minha vida. Todos os desafios da vida tornam-se vencíveis quando eu me lembro de vocês.

Às minhas irmãs, Milena e Cintia, pelo apoio, incentivo e carinho durante toda minha vida.

Aos professores Sérgio de Melo Schneider e Luís Fernando Faina, que enquanto diretores do Centro de Tecnologia da Informação da UFU, me apoiaram e me incentivaram para que eu seguisse adiante no desenvolvimento deste trabalho.

A todos meus amigos e colegas de trabalho do Centro de Tecnologia da Informação da UFU, pelo apoio e incentivo.

Aos meus novos amigos Jean, Walter, Vinícius e Robson, pela paciência, sinceridade, apoio e convivência durante estes anos no laboratório.

A todos os meus amigos, familiares, colegas e professores que diretamente ou indiretamente contribuíram para a realização deste trabalho.

Principalmente à professora Denise Guliato pelo apoio, profissionalismo, paciência, amizade e orientação em todos os momentos da realização deste trabalho. Você é uma pessoa que eu admiro muito, pelo seu bom senso, inteligência, honestidade, caráter e humildade.

*“É sempre mais fácil achar que a culpa é do outro.”
(Raul Seixas)*

Resumo

Nos últimos anos temos visto um grande interesse na recuperação de imagens baseados em conteúdo, com ênfase no desenvolvimento de extratores de características visuais e medidas de similaridade.

Neste trabalho, propomos uma nova abordagem para representar o espaço visual de características, levando em conta as incertezas presentes no processo de extração de característica. A idéia é representar cada dimensão do espaço característica por um conjunto difuso, de acordo com a partição difusa associada a essa dimensão. Devido à representação difusa ser fortemente dependente da partição difusa, propomos também um novo método automático e não supervisionado para obter a partição difusa para cada dimensão do espaço característica com base na técnica de agrupamento Fuzzy C-Means.

Nós testamos a representação difusa, construída a partir de diferentes partições difusas, usando base de dados sintéticas e base de dados reais. A avaliação dos testes indicaram que a representação difusa gerada a partir da partição difusa proposta fornece excelentes resultados. Finalmente, diferentes medidas de similaridades foram aplicadas à representação difusa proposta, indicando que os resultados não são fortemente sensíveis à escolha da medida de similaridade.

Palavras chave: representação por conjuntos difusos, partições difusas, recuperação de imagens por conteúdo, conjuntos difusos.

Abstract

In the recent years we have witnessed great interest in content-based image retrieval with emphasis in the development of visual feature extractors and similarity measures.

In this paper we propose a novel approach to represent the visual feature space, taking into account the uncertainty presents in the extraction feature process. The idea is to represent each dimension of the feature space by a fuzzy set, according to the fuzzy partition associated to this dimension. Because the fuzzy representation is strongly dependent of the fuzzy partition, we also propose a novel automatic unsupervised method to obtain the fuzzy partition for each dimension of the feature space based on Fuzzy C-Means clustering.

We tested the fuzzy representation, constructed from different fuzzy partitions, using synthetic data sets and real data sets. The evaluation of the tests indicated that the fuzzy representation constructed from the proposed fuzzy partition provides excellent results. Finally, different similarity measures were applied to the proposed fuzzy representation, indicating that the results are not strongly sensitive to the choice of the similarity measure.

Keywords: fuzzy set representation, fuzzy partitions, content based image retrieval, fuzzy sets.

Sumário

Lista de Figuras	xi
1 Introdução	13
1.1 Considerações Iniciais	13
1.2 Objetivos	14
1.3 Organização do Trabalho	15
2 Fundamentos Teóricos	17
2.1 Introdução	17
2.2 Conjuntos difusos	17
2.2.1 Operações sobre conjuntos difusos	18
2.2.2 Partições difusas	18
2.3 Método de Agrupamento: Fuzzy C-means	19
2.4 Medidas de avaliação de agrupamentos	21
2.4.1 Coeficiente de silhueta	21
2.4.2 Entropia	22
2.5 Considerações Finais do Capítulo	22
3 Métodos para obtenção de partições difusas	23
3.1 Introdução	23
3.2 Trabalhos correlatos	23
3.3 Detalhamento de alguns métodos	25
3.3.1 Partição Difusa Uniforme	25
3.3.2 Partição Difusa Utilizando o Algoritmo K-means	26
3.3.3 Hierarchical Fuzzy Partitioning (HFP)	27
3.4 Considerações Finais do Capítulo	29
4 Medidas de Similaridade	30
4.1 Introdução	30
4.2 Operadores de similaridades	30
4.3 Considerações Finais do Capítulo	33

5	Representação do Espaço Visual de Característica Utilizando Conjuntos Difusos	34
5.1	Introdução	34
5.2	Uma Nova Representação para o Espaço Visual de Característica	34
5.3	Um exemplo numérico	35
5.4	Considerações Finais do Capítulo	37
6	Método Automático e Não-Supervisionado Para a Geração de Partições Difusas (MANGPD)	39
6.1	Introdução	39
6.2	Método Proposto	39
6.3	Considerações Finais do Capítulo	42
7	Avaliação e Discussão dos Resultados	43
7.1	Introdução	43
7.2	Metodologia utilizada nos experimentos	43
7.3	Base de Dados Sintéticas	44
7.4	Base de Dados Reais	50
7.5	Considerações Finais do Capítulo	60
8	Conclusão	61
	Referências Bibliográficas	62

Lista de Figuras

1.1	Visão geral do processo para representar as características por conjuntos difusos.	15
2.1	Funções de pertinência.	18
2.2	Partição difusa do universo idade.	19
3.1	Partição difusa uniforme da dimensão d_1 . O domínio de d_1 é $[0, 10]$	26
3.2	Partição obtida utilizando o algoritmo K-means para a dimensão d_1 . O domínio de d_1 é $[0, 10]$	27
3.3	Junção dos conjuntos difusos 2 e 3 resultando em $2'$, $1 \Rightarrow 1'$, $4 \Rightarrow 3'$. Os c^i são os centros dos conjuntos difusos.	28
5.1	Um exemplo numérico simples. O domínio de d_1 é o intervalo $[0,5]$ e o domínio de d_2 é o intervalo $[0,10]$	35
5.2	Partição difusa para cada dimensão do espaço característica mostrado na Figura 5.1.	36
6.1	Resultado do FCM em uma dimensão do espaço de característica, no qual foram encontrados três agrupamentos. O valor 6 possui os graus de pertinência 0.18 em relação ao primeiro agrupamento, 0.72 em relação ao segundo agrupamento e 0.10 em relação ao terceiro agrupamento.	40
6.2	Pontos para identificação das retas para a geração de duas funções de pertinência adjacentes, tendo como base um resultado do FCM.	40
6.3	Partição difusa obtida pelo método proposto. A linha sólida indica o resultado do algoritmo FCM e a linha tracejada indica a funções de pertinência encontradas pelo método proposto.	41
7.1	Base de dados BDSintetica01.	44
7.2	Base de dados BDSintetica02.	44
7.3	Base de dados BDSintetica03.	45
7.4	Experimentos usando a base de dados sintética BDSintetica01.	46
7.5	Experimentos usando a base de dados sintética BDSintetica02.	47

7.6	Experimentos usando a base de dados sintética BDSintetica03.	48
7.7	Experimentos usando a base de dados IRIS.	53
7.8	Experimentos usando a base de dados WDBC.	54
7.9	Experimentos usando a base de dados BD111.	55
7.10	Experimentos usando a base de dados BDFACES.	56
7.11	Experimentos usando a base de dados BDSHAPES.	57

Capítulo 1

Introdução

1.1 Considerações Iniciais

Nos dias de hoje é notável o avanço de hardware dos computadores, aumento do número de redes conectadas a Internet, e a capacidade de armazenamento de dados cada vez maior, propiciando um grande crescimento da informação digital. Junto a este crescimento de informações há também o surgimento de novas aplicações para tratar estes dados, os quais não estão limitados a apenas dados textuais ou numéricos, existindo também vídeos, dados de áudio e imagens.

O crescimento no número de base de dados de imagens em vários campos como imagens médicas, sistemas de informações geográficas, bibliotecas multimídias e outros, vem despertando grande interesse para o desenvolvimento de técnicas para Recuperação de Imagens Baseadas em Conteúdo (CBIR). Nos últimos anos, foram apresentados na literatura acadêmica diversos trabalhos relacionados a CBIR [Datta et al. 2008], [Flickner et al. 1995], [Smith e Chang 1997] e [Veltkamp e Tanase 2002].

A comparação entre as imagens é uma operação fundamental para os sistemas de recuperação de imagens baseada em conteúdo (CBIR). Os métodos descritos na literatura para realizar tal comparação são geralmente divididos em dois grupos: o primeiro é baseado em identificar a similaridade entre duas imagens considerando a análise dos pixels das respectivas imagens e o segundo método é baseado em características visuais extraídas das imagens.

Após um processo de extração de características, geralmente baseado em cor, forma ou textura, uma imagem é representada por um vetor de característica, no qual cada elemento do vetor é representado por um único valor numérico e está associado a uma característica visual da imagem extraída. Essa nova representação da imagem visa caracterizar uma imagem ou uma classe de imagens e ser invariante às transformações da imagem.

Tendo como base a representação numérica das imagens, ou seja, os vetores de características, o grau de similaridade entre duas imagens é então derivado pelo cálculo da (di)

similaridade entre seus respectivos vetores de características [Hjaltason e Samet 2003]. A (di) similaridade calculada é fortemente dependente da escolha do conjunto de características visuais das imagens que serão extraídos [Guliato et al. 2008], [Guliato et al. 2009], [Datta et al. 2008], [Choras et al. 2007] da representação destas características, bem como da escolha de medidas de similaridade.

De acordo com [Candan et al. 2000], a obtenção dos vetores de características visuais de uma imagem incorpora incertezas resultantes tanto do processo de aquisição da imagem quanto do processo de segmentação e extração de características.

De um modo geral, as medidas de similaridade são aplicadas a vetores de características com os valores normalizados, considerando-se ou não, a dependência entre os elementos e a incerteza presente na representação das características. Neste trabalho é proposta uma nova representação das características, em que cada característica é representada por um conjunto difuso.

1.2 Objetivos

O objetivo deste trabalho é propor uma nova abordagem para representar o espaço visual de características, levando-se em conta as incertezas inerentes nos processos de extração das mesmas. A idéia é representar cada característica por um conjunto difuso, de acordo com a partição difusa associada a esta característica.

Como a representação difusa de uma dada característica é fortemente dependente da partição difusa associada ao domínio desta característica, também é proposto, neste trabalho, um método automático e não supervisionado para obtenção da partição difusa para cada dimensão do espaço de características.

Tanto a nova representação de características como o método para geração automática de partições difusas serão avaliados em bases de dados sintéticas e reais.

A Figura 1.1 apresenta uma visão geral do processo para representar as características por conjuntos difusos. Os módulos hachurados foram desenvolvidos neste trabalho.

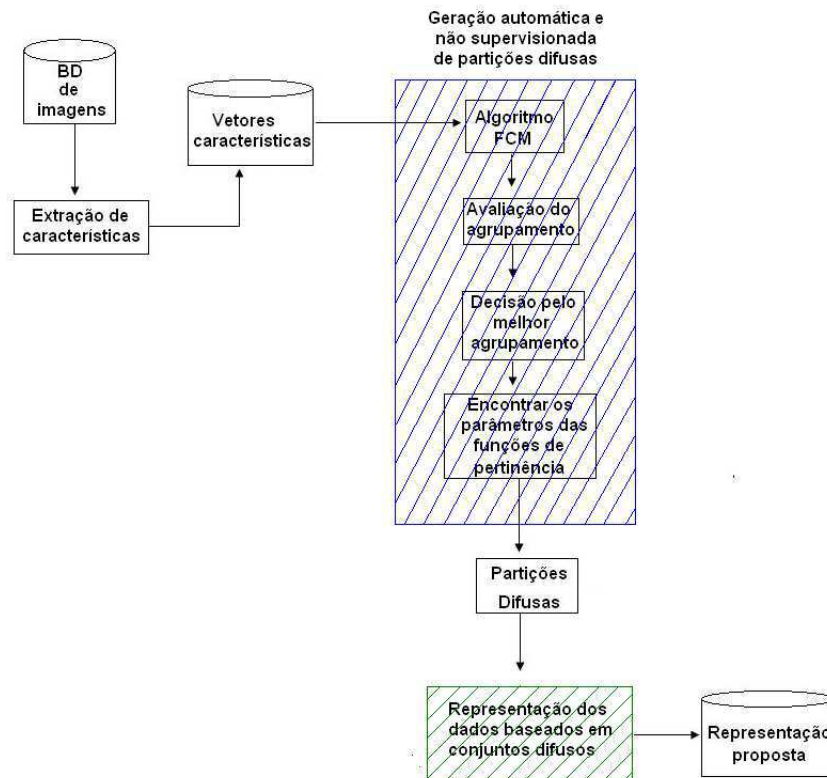


Figura 1.1: Visão geral do processo para representar as características por conjuntos difusos.

1.3 Organização do Trabalho

Este trabalho está organizado da seguinte forma:

- No capítulo 2 são mostrados os conceitos teóricos que fundamentaram o trabalho, tais como a definição de conjuntos difusos, partições difusas e técnicas de agrupamento.
- No capítulo 3 são descritos alguns métodos para obtenção de partições difusas encontrados na literatura. Alguns destes métodos são relatados de forma mais detalhada.
- No capítulo 4 são descritos diversos operadores de similaridade para conjuntos difusos presentes na literatura.
- No capítulo 5 é proposta uma nova forma de representar os vetores de característica, passando a representar cada característica por um conjunto difuso (módulo hachurado de verde na Figura 1.1).
- No capítulo 6 é proposto um novo método para obtenção da partição difusa (módulo hachurado de azul na Figura 1.1).

- No capítulo 7 é apresentada a análise dos resultados, comparando a representação proposta em relação à representação normalizada e também comparando o método para geração de partição difusa proposto com outros métodos existentes.
- No capítulo 8 são apresentadas as conclusões do trabalho e trabalhos futuros.

Capítulo 2

Fundamentos Teóricos

2.1 Introdução

O objetivo deste capítulo é apresentar os conceitos teóricos que deram suporte a este trabalho, dentre os quais estão: os conjuntos difusos, técnicas de agrupamentos e análise da qualidade de agrupamentos.

2.2 Conjuntos difusos

A teoria dos conjuntos difusos foi proposto por Zadeh [Zadeh 1965] e tem como objetivo modelar matematicamente aspectos relacionados à incerteza e vagueza das informações.

Os conjuntos difusos são uma generalização da teoria dos conjuntos clássicos, e possibilitam uma modelagem mais adequada a termos da linguagem natural que não se limitam à classificação "verdadeiro ou falso" da lógica clássica. Tradicionalmente, uma proposição na lógica clássica possui dois extremos: completamente verdadeiro ou completamente falso. Dentro da lógica difusa, uma premissa varia em grau de verdade de 0 a 1, o que possibilita a mesma ser parcialmente verdadeira e parcialmente falsa.

Após o surgimento dos conjuntos difusos, os mesmos tem sido largamente utilizados em sistemas de controles difusos [Mendel 1995], métodos para segmentação de imagens [Pal e Pal 1993], [Faria et al. 2005], [Gulato et al. 1998] e [Tobias e Seara 2002], métodos de classificação [Gulato e Santos 2009], [Kuncheva 2000] e [Luukka 2007].

Definição 1. Um conjunto difuso A definido no universo de discurso X é caracterizado por uma função de pertinência μ_A , que mapeia os elementos de X para o intervalo $[0,1]$. Uma definição mais formal de um conjunto difuso A é dada por:

$$A = \{(x, \mu_A(x)) | x \in X\}, \quad \mu_A(x) : X \longrightarrow [0, 1] \quad (2.1)$$

A função de pertinência normalmente é criada para expressar algum conceito impreciso pela utilização de uma variável linguística. Entre as formas de representação das funções de pertinência, as mais utilizadas são a forma triangular e trapezoidal mostradas na Figura 2.1.

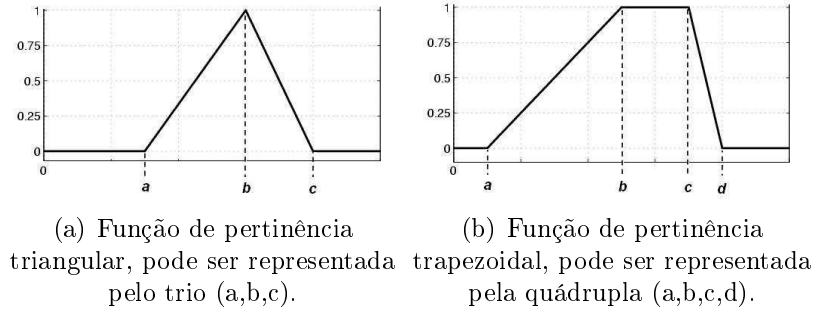


Figura 2.1: Funções de pertinência.

2.2.1 Operações sobre conjuntos difusos

A conjunção e a disjunção de conjuntos difusos são obtidas utilizando-se os conceitos de T -norma e T -conorma, respectivamente. As T -normas constituem uma família de operadores $T : [0, 1] \times [0, 1] \rightarrow [0, 1]$ tais que, para quaisquer $a, b, c, d \in [0, 1]$, T satisfaz as seguintes propriedades:

- Comutatividade: $T(a, b) = T(b, a)$
- Associatividade: $T(a, T(b, c)) = T(T(a, b), c)$
- Monotonicidade: $T(a, b) \leq T(c, d)$ se $a \leq c$ e $b \leq d$
- Elemento neutro: $T(a, 1) = a$

As T -conormas, da mesma forma, são funções $S : [0, 1] \times [0, 1] \rightarrow [0, 1]$ com propriedades comutativa, associativa e monotônica. A diferença é que neste caso o elemento neutro é igual a 0, ou seja, $S(a, 0) = a$.

A família de operadores T -conorma representa a união entre dois conjuntos difusos, um exemplo é o operador $\max()$. A família T -norma representa a intersecção entre dois conjuntos difusos, um exemplo é o operador $\min()$ [Klir e Yuan 1995].

2.2.2 Partições difusas

Uma partição difusa é caracterizada por um ou mais conjuntos difusos definido em um mesmo universo.

Segundo [Klir e Yuan 1995], dado um universo do discurso $X = \{x_1, x_2, \dots, x_n\}$, sendo $CD(X)$ a família dos conjuntos difusos deste universo, $CD(X) = \{A_1, A_2, \dots, A_c\}$ e A_i

é um conjunto difuso. Uma partição difusa deste universo deve satisfazer as seguintes condições:

- Para todo k , com k sendo inteiro e $1 \leq k \leq n$, $\sum_{i=1}^c \mu_{A_i}(x_k) = 1$;
- Para todo i , com i sendo inteiro e $1 \leq i \leq c$, $0 < \sum_{k=1}^n \mu_{A_i}(x_k) < n$.

A Figura 2.2 mostra um exemplo de uma partição difusa referente o universo idade. No exemplo, a partição difusa é composta de três conjuntos difusos.

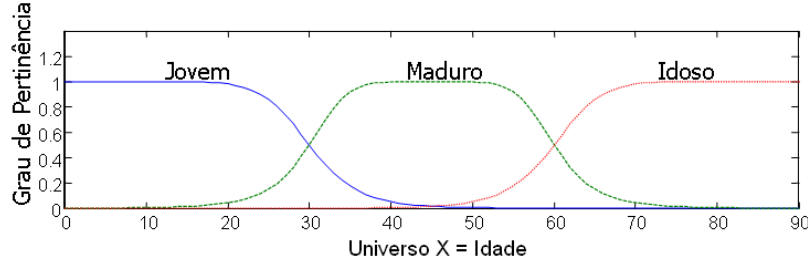


Figura 2.2: Partição difusa do universo idade.

2.3 Método de Agrupamento: Fuzzy C-means

Os métodos de agrupamento podem ser divididos em particionais e hierárquicos. Nas técnicas particionais clássicas, os agrupamentos são disjuntos e não possuem dependências entre si e nas técnicas hierárquicas os agrupamentos possuem subagrupamentos como em uma estrutura de árvore em que o nó interno da árvore é a união de seus filhos ([Steinbach et al. 2006]).

Entre as técnicas clássicas de agrupamento particionais, citamos o método K-means [MacQueen 1967] como um dos métodos mais populares, que tem como objetivo particionar os dados em K agrupamentos mutuamente exclusivos. O Fuzzy C-means (FCM) é uma extensão do K-means e permite que um dado objeto possa pertencer a mais de um agrupamento. Este método foi desenvolvido por Dunn [Dunn 1973] e aperfeiçoado por Bezdek [Bezdek 1981]. O FCM começa com uma suposição inicial sobre os centros de cada agrupamento. Cada objeto da base de dados é atribuído a cada agrupamento com um grau de pertinência. O valor do grau de pertinência de um objeto a um determinado agrupamento está no intervalo $[0, 1]$. Em um procedimento iterativo, os centros de cada agrupamento e também os graus de pertinência de cada objeto relacionado a cada agrupamento são atualizados. O objetivo deste procedimento é minimizar a função objetivo mostrada na Equação 2.2.

$$J_m = \sum_{i=1}^n \sum_{k=1}^K \mu_{ik}^m \|x_i - c_k\|^2 \quad (2.2)$$

em que m é o coeficiente difuso, n é o número total de objetos do conjunto de dados, K é o número de agrupamentos, x_i é numérico e representa o objeto i – *esimo* desse conjunto de dados, c_k é o centro do agrupamento k e μ_{ik} é grau de pertinência do objeto x_i em relação ao agrupamento k .

A cada iteração, o centro do agrupamento k é obtido de acordo com a Equação 2.3.

$$c_k = \frac{\sum_{i=1}^n \mu_{ik} \cdot x_i}{\sum_{i=1}^n \mu_{ik}} \quad (2.3)$$

Uma das restrições presentes é o fato de que apesar de um objeto x_i qualquer poder pertencer a mais de um agrupamento com os graus de pertinência em relação a cada agrupamento, a soma de todos estes graus de pertinência deve ser igual a 1. Considerando K agrupamentos e μ_{ik} o grau de pertinência do objeto x_i em relação ao agrupamento k , temos:

$$\sum_{k=1}^K \mu_{ik} = 1 \quad (2.4)$$

Para executar o algoritmo é necessário definir o critério de distância a ser utilizado, o número de agrupamentos K , o valor de m e um valor ϵ como critério de parada. Segundo [Hathaway e Bezdek 2001], uma boa escolha para m , é $m = 2$, uma vez que é possível verificar alguma similaridade entre os agrupamentos mantendo ainda as suas características próprias.

O Algoritmo do FCM é dado abaixo:

Algorithm 1 Algoritmo FCM

- 1: Inicializar:
 K (número de agrupamentos),
 ϵ (critério de parada),
 m (coeficiente difuso),
 $U^{(0)}$ (matriz inicial com os graus de pertinência, dimensão da matriz n, K) e
 $t = 0$ (iteração inicial)
- 2: Calcular os centros dos K agrupamentos $(c_1^{(t)}, c_2^{(t)}, \dots, c_K^{(t)})$, sendo t a iteração corrente
- 3: Atualizar a matriz $U^{(t+1)}$ segundo a Equação 2.5:

$$\mu_{ik}^{(t+1)} = \frac{1}{\sum_{j=1}^n \left(\frac{\|x_i - c_k^{(t)}\|^2}{\|x_i - c_j^{(t)}\|^2} \right)^{\frac{1}{m-1}}} \quad (2.5)$$

onde $\mu_{ik}^{(t+1)}$ é o grau de pertinência do objeto i em relação ao agrupamento k na iteração $t + 1$ e x_i é o objeto i do conjunto de dados.

- 4: Se $\|U^{(t+1)} - U^{(t)}\| < \epsilon$ ou (um número de iterações pré-determinado é executado) então fim do algoritmo, senão, $t = t + 1$ e voltar ao passo 2.
-

2.4 Medidas de avaliação de agrupamentos

A avaliação da qualidade dos agrupamentos é uma das tarefas mais difíceis no processo de agrupamento. É pela avaliação da qualidade dos agrupamentos que procura-se responder questões como "Qual é o número ideal de agrupamentos?" ou "Considerando dois conjuntos de agrupamentos obtidos para os mesmos dados, qual é o melhor?".

No desenvolvimento de medidas numéricas para avaliar os agrupamentos há alguns pontos que devem ser considerados, como a interpretabilidade da medida e também a complexidade para a obtenção da mesma.

As medidas de avaliação de agrupamentos são classificadas em:

- Supervisionadas: são medidas que medem a qualidade de um agrupamento resultante utilizando informações externas aos dados, como por exemplo, dados de testes contendo um atributo representando classe.

- Não-supervisionada: medem a qualidade do agrupamento resultante sem utilizar medidas externas aos dados. As medidas não supervisionadas são frequentemente divididas em medidas de coesão e separação. As medidas de coesão avaliam o quanto os objetos dentro de um agrupamento estão próximo ao centro deste agrupamento, e as medidas de separação avaliam o quanto os centros dos agrupamentos estão bem distantes entre si.

2.4.1 Coeficiente de silhueta

O coeficiente de silhueta é uma medida não supervisionada que combina coesão e separação [Steinbach et al. 2006]. Os passos seguintes mostram como calcular o coeficiente de silhueta para um objeto individualmente:

- 1 - Para o i^{th} objeto, calcula-se a_i , sendo a_i igual à distância média do objeto i a todos os outros objetos pertencentes ao seu agrupamento.

- 2 - Para o i^{th} objeto, calcula-se b_i , sendo b_i igual à distância média do objeto i a todos os outros objetos pertencentes ao agrupamento B (agrupamento vizinho ao agrupamento do objeto i). O agrupamento vizinho é um tipo de segundo melhor agrupamento para o objeto i . Para encontrar o agrupamento vizinho, considere qualquer agrupamento C que não contenha o objeto i e também $d(i, C)$ como sendo a distância média do objeto i a todos os objetos de C . Levando em conta todos os agrupamentos que não contenham o objeto i , o agrupamento vizinho ao agrupamento do objeto i é o agrupamento que $d(i, C)$ apresenta o menor valor.

- 3 - O coeficiente de silhueta para o objeto i é dado por $s_i = (b_i - a_i) / \max(a_i, b_i)$

O coeficiente de silhueta varia de -1 a 1. Os valores ideais são valores positivos e a_i bem próximo de zero. A interpretação dos valores é feita da seguinte maneira: quanto mais próximo de 1, melhor o agrupamento resultante.

O coeficiente de silhueta de um determinado agrupamento é obtido pela média dos coeficientes de silhueta dos objetos pertencentes a este agrupamento, e o coeficiente de

silhueta da partição resultante é obtido pela média dos coeficientes de silhuetas de todos os objetos.

No caso do agrupamento resultante produzido pelo Fuzzy C-Means, em que um mesmo objeto pertence a mais de um agrupamento com diferentes graus de pertinência, é considerado para efeito do cálculo do coeficiente de silhueta, aquele agrupamento ao qual o objeto pertence com maior grau de pertinência [Campello e Hruschka 2006].

2.4.2 Entropia

A entropia é uma medida supervisionada que utiliza o atributo classe dos objetos. Esta medida avalia o grau em que cada agrupamento é composto por elementos de uma única classe [Steinbach et al. 2006].

A entropia de um agrupamento c é dada por:

$$e_c = - \sum_{j=1}^L \frac{m_{cj}}{m_c} \log_2 \frac{m_{cj}}{m_c} \quad (2.6)$$

onde

L é o número de classes

m_{cj} é o número de elementos do agrupamento c que pertencem a classe j

m_c é o número de elementos do agrupamento c .

A entropia total do agrupamento resultante é calculada pela média ponderada das entropias de cada agrupamento conforme Equação 2.7.

$$e = \sum_{c=1}^K \frac{m_c}{m} e_c \quad (2.7)$$

onde K é o número de agrupamentos e m é o número total de elementos.

Os valores da entropia variam de 0 a $\log_2 L$. O caso ideal de um agrupamento seria quando ele é composto por elementos de uma única classe ($e = 0$). O pior caso de um agrupamento é quando ele é composto por elementos de todas as classes possíveis, e o número de elementos de cada classe é exatamente o mesmo ($e = \log_2 L$, ou seja, a distribuição é uniforme). A entropia normalizada é calculada por $e_n = e / \log_2 L$.

2.5 Considerações Finais do Capítulo

Após o surgimento da teoria dos conjuntos difusos, que possibilitou a modelagem matemática de incertezas, a mesma teve uma grande aplicabilidade em diversas áreas.

Neste capítulo foram apresentados alguns dos conceitos teóricos que sustentaram o desenvolvimento deste trabalho. No próximo capítulo serão descritos algumas das técnicas relacionadas na literatura para a definição de partições difusas.

Capítulo 3

Métodos para obtenção de partições difusas

3.1 Introdução

As partições difusas são muito utilizadas em sistemas de classificação difusos, sistemas de controle difusos e também em sistemas especialistas difusos. Uma tarefa crítica no desenvolvimento destes sistemas é a pré-definição das funções de pertinências para mapear dados numéricos em termos de variáveis linguísticas, possibilitando o raciocínio em termos da lógica difusa. Como exemplo, citamos os sistemas de classificação baseados em regras difusas utilizando as partições difusas para mapear características em termos linguísticos para gerar regras do tipo IF-THEN.

Neste capítulo são destacados diversos trabalhos já realizados que utilizaram algum método para a definição de partições difusas. Também é dado um maior detalhamento em nas das estratégias para obtenção das partições difusas que serão usadas posteriormente.

3.2 Trabalhos correlatos

A definição das funções de pertinência difusa é uma tarefa de grande importância em sistemas que necessitam do mapeamento dos dados numéricos para termos linguísticos. No trabalho [Medasani et al. 1998] foram apresentadas algumas das idéias dos métodos para a geração automática de partições difusas visando o reconhecimento de padrões, entre as quais estão: métodos baseados na percepção, métodos heurísticos, métodos baseados em histogramas e transformação de distribuição de probabilidade em distribuição de possibilidade.

O conjunto das funções de pertinência para cada dimensão do espaço do domínio da aplicação formam a partição difusa daquela dimensão, sendo esta partição crucial para o desempenho dos sistemas que a utilizam.

Em [Ishibuchi et al. 1992] foi proposto um sistema de classificação baseado em regras do tipo IF-THEN. A fim de obter as regras, os autores utilizaram partições difusas obtidas a partir dos dados numéricos. A partição difusa do espaço de dados foi obtida dividindo o espaço de dados uniformemente (cada dimensão dividida em k conjuntos difusos).

Uma outra técnica para geração automática da partição difusa foi apresentada em [Kbir et al. 2000]. Os autores propuseram um método supervisionado e hierárquico baseado na decomposição do espaço de dados em 2^N subespaços. A decomposição é controlada por dois parâmetros pré-definidos: o grau de certeza das regras difusas para cada subespaço e o nível máximo de decomposição. Para cada subespaço, as regras difusas são geradas e o grau de certeza é calculado utilizando a base de dados. Apenas funções de pertinência triangulares simétricas são geradas, possuindo uma grande área de sobreposição entre funções adjacentes.

Nos trabalhos [Hong e Lee 1996] e [Hong e Chen 1999] foi desenvolvido um sistema de classificação baseado em regras difusas do tipo IF-THEN e também um método supervisionado para encontrar as partições difusas. Para geração das partições difusas, foi utilizada uma tabela de decisão multidimensional construída inicialmente de acordo com a base de dados de treinamento. Com a aplicação do algoritmo proposto, esta tabela sofria junções entre as linhas e/ou colunas respeitando algumas regras pré-definidas. Ao final do algoritmo, a tabela de decisão indicava não só as regras IF-THEN como também as partições difusas encontradas.

Técnicas de agrupamento também são utilizadas para a definição das partições. Em [Guillaume et al. 2001], foi utilizado o algoritmo K-means sobre cada dimensão do espaço de dados. Os centróides dos agrupamentos foram considerados como os centros das funções de pertinência definidas como funções triangulares, deixando para o usuário a escolha do tamanho da partição difusa. Em [Chiu 1994], foi proposto um método para estimar os centros dos agrupamentos. A idéia desse método é atribuir um valor numérico a cada objeto do espaço de dados, que representa o potencial de cada objeto ser o centro de um agrupamento, sendo que um objeto com muitos vizinhos próximos tem seu valor numérico alto. O objeto com maior potencial é escolhido como o centro do primeiro agrupamento e os valores numéricos dos outros objetos são reduzidos de acordo com a distância ao centro do agrupamento já definido. Para o próximo centro de agrupamento, é escolhido entre os objetos restantes o objeto com maior potencial, recalculando os valores numéricos dos outros objetos. Este procedimento de escolha dos centros de agrupamento e consequentemente recálculo dos potenciais dos objetos é repetido até um critério de parada. Ao final do método tem-se os centros dos agrupamentos definidos, os quais são utilizados como os centros das funções de pertinência triangulares. Nota-se que o número de funções de pertinência é o mesmo em todas as dimensões.

No trabalho [Guillaume e Charnomordic 2004], foi proposto o método chamado Hierarchical Fuzzy Partitioning (HFP), o qual é inspirado em agrupamento hierárquico e

conjuntos difusos. O HFP é um método que trabalha com as dimensões do espaço de dados de forma independente. O procedimento tem como base uma partição inicial composta por n funções de pertinência e vai realizando junções de funções adjacentes de acordo com um critério que indica qual a melhor junção a fazer. A cada iteração, diminui em um o número de funções de pertinência. No final do processamento, tem-se as melhores partições de tamanho n (partição inicial) até 2 conjuntos difusos. Para a definição da melhor partição, os autores propõem um novo índice de avaliação.

Também há trabalhos descritos na literatura para a definição das partições difusas que empregam os conceitos de algoritmos genéticos, como os trabalhos de [Murata et al. 1998] e [Ishibuchi e Nakashima 2000].

3.3 Detalhamento de alguns métodos

Na Seção anterior foram citados alguns dos trabalhos que propuseram algum método para a obtenção da partição difusa. Os métodos baseados em partição difusa uniforme, partição difusa obtida utilizando k-means e Hierarquical Fuzzy Partition foram utilizados para a avaliação do método de obtenção de partição difusa proposto neste trabalho. Por este motivo, faz-se necessário um maior detalhamento destes métodos.

3.3.1 Partição Difusa Uniforme

Neste método, a obtenção da partição difusa é feita pela divisão uniforme do espaço de características em K partes de mesmo tamanho.

Primeiramente, cada dimensão é analisada de forma independente. Considerando uma dimensão, faz-se a divisão da mesma em K partes iguais, ou seja, dado o domínio $[Menor\ valor, Maior\ valor]$ para a dimensão, o tamanho de cada parte é calculado por $(Maior\ valor - Menor\ valor)/K$ e para cada parte calcula-se o centro da mesma.

A partir da divisão da dimensão, deve-se encontrar os conjuntos difusos que farão parte da partição difusa. As funções de pertinência são da forma triangular, com exceção do primeiro e último conjunto difuso, que estão nas extremidades da dimensão e possuem funções de pertinência da forma semi trapezoidal. Os centros dos intervalos são utilizados como os centros das funções de pertinência. As funções de pertinência com forma triangular ou semi trapezoidal podem ser representadas por três parâmetros (a , b , c).

Podemos representar a primeira função de pertinência difusa pelos parâmetros $(a_{primeira}, b_{primeira}, c_{primeira})$ e a última função de pertinência difusa por $(a_{ultima}, b_{ultima}, c_{ultima})$, sendo: $a_{primeira}$ o menor valor do domínio; $b_{primeira}$ o centro do primeiro intervalo; $c_{primeira}$ o centro do segundo intervalo; a_{ultima} o centro do penúltimo intervalo; b_{ultima} o centro do último intervalo; c_{ultima} o maior valor do domínio.

As funções de pertinência difusas com formas triangulares são representadas pelos

parâmetros (a_i, b_i, c_i) , em que : a_i o centro do intervalo $i - 1$; b_i o centro do intervalo i ; c_i o centro do intervalo $i + 1$, $i > 1$ e $i < K$.

Na Figura 3.1 é mostrada a partição uniforme obtida para uma dada dimensão $d1$ pertencendo ao domínio $[0, 10]$ e considerando que a mesma é composta de 5 conjuntos difusos ($K = 5$). Note que os valores dos centros das funções de pertinência são 1, 3, 5, 7 e 9.

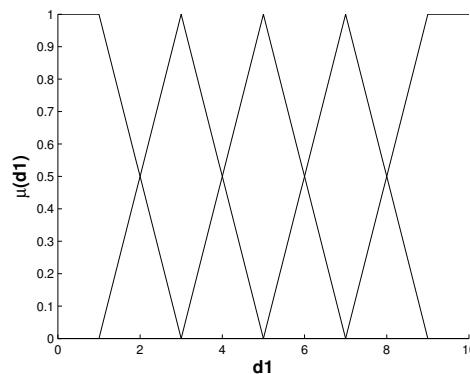


Figura 3.1: Partição difusa uniforme da dimensão $d1$. O domínio de $d1$ é $[0, 10]$.

Pode-se observar que os conjuntos difusos cruzam-se no ponto em que os seus valores de pertinência são iguais a 0,5.

3.3.2 Partição Difusa Utilizando o Algoritmo K-means

Considerando cada dimensão de forma independente, uma outra estratégia para obtenção da partição difusa é utilizando o algoritmo de agrupamento K-means [MacQueen 1967]. No algoritmo K-means é fornecido como parâmetro o número de agrupamentos K . O método K-means garante que os elementos de um mesmo grupo estão próximos (de acordo com um critério dado) e os elementos de grupos distintos estão longe de acordo com este mesmo critério.

O objetivo de utilizar-se do algoritmo de agrupamento K-means é encontrar os centróides dos agrupamentos. Tendo encontrado os centróides, encontra-se os conjuntos difusos utilizando-se dos centróides como os centros das funções de pertinência. As funções de pertinência localizadas nas extremidades possuem forma semi trapezoidal e as demais são da forma triangular, e são obtidas da mesma forma que o método de geração das partições difusas apresentado na subseção anterior.

Considerando uma dada dimensão $d1$ pertencendo ao domínio $[0, 10]$ e que os centróides resultantes do K-means para K igual a 4 sejam 2, 5,1, 6,5 e 9, a partição obtida é mostrada na Figura 3.2.

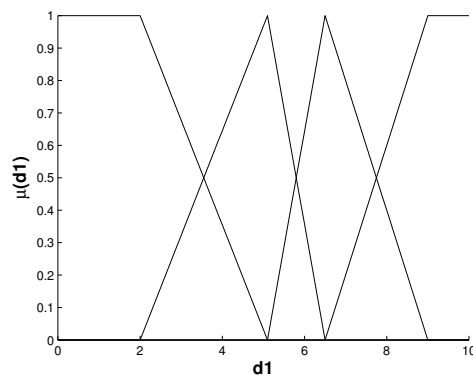


Figura 3.2: Partição obtida utilizando o algoritmo K-means para a dimensão $d1$. O domínio de $d1$ é $[0, 10]$.

Neste método os conjuntos difusos também cruzam-se no ponto em que os seus valores de pertinência são iguais a 0,5. A diferença com relação ao método baseado em partições uniformes, é que cada dimensão do espaço característica é dividida de acordo com a distribuição dos dados naquela dimensão.

Ao utilizar os centróides como sendo os centros das funções de pertinência, este método assume como uma limitação o fato de gerar funções de pertinência apenas com a forma triangular.

3.3.3 Hierarchical Fuzzy Partitioning (HFP)

O HFP foi proposto em [Guillaume e Charnomordic 2004]. É um método que trata cada dimensão de forma independente, gerando uma família de partições em cada dimensão. O HFP é baseado em agrupamentos hierárquicos aglomerativos, ou seja, parte de um número de agrupamentos inicial igual ao número de elementos e vai executando a junção de agrupamentos até que no final encontre um único agrupamento.

A partição difusa inicial é determinada pela escolha do número de conjuntos difusos iniciais, sendo este um dos parâmetros do método, respeitando sempre o limite de que o número de conjuntos difusos iniciais deve ser menor ou igual ao número total de elementos da base de dados.

A família de partições difusas é obtida usando um procedimento recursivo, no qual a cada passo, a partição resultante possui o número de conjuntos difusos do passo anterior diminuído de um, devido a uma junção de conjuntos difusos adjacentes que atendem um determinado critério. A junção é restrita aos conjuntos difusos adjacentes, e afeta no máximo quatro conjuntos. Na figura 3.3 é mostrada uma possível junção entre os conjuntos difusos 2 e 3.

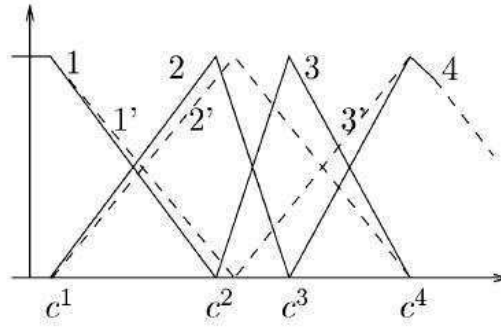


Figura 3.3: Junção dos conjuntos difusos 2 e 3 resultando em 2', $1 \Rightarrow 1'$, $4 \Rightarrow 3'$. Os c^i são os centros dos conjuntos difusos.

O algoritmo para obter a família de partições difusas é apresentado a seguir

Algorithm 2 Algoritmo HFP

```

1: base partition =  $FP^m$  = initial partition of size  $M_j$ , set  $m = M_j$ 
2: while  $m > 2$  do
3:   evaluate  $D_m$ ;  $s = 1$ 
4:   while  $s \leq m - 1$  do
5:     merge fuzzy sets  $s$  and  $s + 1$ , modify neighboring fuzzy sets
6:     evaluate  $D_{m-1}^s$ 
7:     restore base partition
8:      $s = s + 1$ 
9:   end while
10:  select and store the partition  $FP^{m-1}$  for which:
11:     $D_{m-1} = \operatorname{argmax}(D_{m-1}^s)$ 
12:    base partition =  $FP^{m-1}$ ;  $m = m - 1$ ;
13: end while

```

O critério D_m representa a soma das distâncias entre os pares de pontos p e q , levando-se em conta todos os pontos de dados. A melhor junção é obtida quando minimiza a variação de D_m .

D_m é dado por:

$$D_m = \frac{1}{N(N-1)} \sum_{q,r=1,2,\dots,N,q \neq r} d(q,r)$$

em que

$$d(q,r) = \frac{1}{\sum_{f=1}^m \mu_q^f} \sum_{f=1}^m \left[\mu_q^f \frac{1}{\sum_{g=1}^m \mu_r^g} \sum_{g=1}^m [\mu_r^g d_{f,g}(q,r)] \right]$$

$d_{f,g}(q,r)$ pode ser classificado com uma distância interna ou externa. A distância interna é quando os dois pontos p e q possuem algum grau de pertinência em um mesmo conjunto difuso e distância externa é quando os dois pontos p e q não possuem algum grau de pertinência em um mesmo conjunto difuso. Para distância interna $d_{f,g}(q,r) = |\mu_q^f - \mu_r^f|$, sendo f e g o mesmo conjunto difuso. Para distância externa $d_{f,g}(q,r) = |\mu_q^f - \mu_r^g| + \sqrt{(c_f - c_g)^2} + D_c$, sendo c_f e c_g os centros dos conjuntos difusos e D_c um fator de correção constante para garantir que nenhuma distância externa seja menor que uma distância interna.

Uma vez obtida a família de partições difusas, faz-se necessário escolher a melhor partição. Para selecionar a melhor partição, os autores propuseram um índice baseado na homogeneidade das densidades dos conjunto difusos que é dada por:

$$\sigma^{FP} = \sqrt{1/m \sum_{f=1}^m (d_f - dm)^2} \quad (3.1)$$

em que

d_f é a densidade do conjunto difuso f . A densidade do conjunto difuso é igual à razão do seu peso (ou cardinalidade) pela sua área e dm é a densidade média da partição difusa.

O índice σ^{FP} não tem muito significado para os primeiros passos do algoritmo quando as partições difusas possuem muitos conjuntos difusos. Ele é útil para as últimas etapas, quando as partições difusas são formadas por sete conjuntos difusos ou menos. Do ponto de vista da homogeneidade a melhor partição é aquela para o qual este valor atinge um mínimo.

Assim como nos métodos detalhados anteriormente, os conjuntos difusos das partições são da forma triangular, com exceção do primeiro e último conjunto, que estão nas extremidades da dimensão e são da forma semi trapezoidal. E também, dois conjuntos difusos adjacentes cruzam-se no valor de suas funções de pertinência igual 0,5.

Este método produz funções de pertinência difusas com grande área de sobreposição e possui alto custo computacional.

3.4 Considerações Finais do Capítulo

A pré-definição das partições difusas é uma das tarefas comumente realizada em sistemas baseados na lógica difusa. Neste capítulo foram relatados diversos trabalhos que utilizaram algum método automático para a definição das partições difusas, e também foram apresentados com mais detalhes alguns destes métodos.

Os métodos Partição Difusa Uniforme, Partição Difusa Utilizando o Algoritmo K-means e HFP se limitam a gerar funções de pertinência apenas na forma triangular, e também as funções de pertinência geradas por eles possuem uma alta sobreposição. Nós apresentamos no Capítulo 6 um novo método automático para obter a partição difusa que possibilita gerar funções de pertinência com diferentes formas e com uma menor área de sobreposição entre as funções de pertinência.

Capítulo 4

Medidas de Similaridade

4.1 Introdução

As medidas de similaridade são utilizadas em sistemas computacionais para calcular o quanto dois objetos são semelhantes. Sistemas de recuperação por conteúdo e sistemas de classificação são bons exemplos de sistemas que geralmente utilizam medidas de similaridade.

Ao pensar em similaridade entre objetos, é natural pensar em duas imagens ou mesmo dois elementos representando classes. Mas medidas de similaridades também podem ser aplicadas para determinar o quanto dois conjuntos difusos são similares.

Neste capítulo são apresentados alguns operadores de similaridade para conjuntos difusos descritos na literatura.

4.2 Operadores de similaridades

Para avaliar a similaridade entre dois objetos faz-se necessário a determinação de um valor numérico que expresse o quanto estes objetos são similares.

Utilizando-se vetores de características para representar os objetos, pode-se obter o valor da similaridade pela aplicação de medidas de distância entre os vetores destes objetos. Por se tratar de medidas de distâncias, o valor encontrado representa o quanto os objetos estão distantes ou são dissimilares, sendo que quanto maior o valor, menor será a similaridade entre eles, e vice versa.

A definição de um espaço métrico é dada por: $M\{C, d()\}$, onde C é o conjunto dos elementos do domínio e $d()$ é uma função de distância. Considerando os elementos e_1 , e_2 e e_3 pertencentes ao domínio C , uma função de distância para o espaço métrico $M\{C, d()\}$ deve satisfazer as seguintes propriedades.

- Não negatividade: $d(e_1, e_2) \geq 0$;
- Simetria: $d(e_1, e_2) = d(e_2, e_1)$;

- Desigualdade triangular: $d(e_1, e_2) \leq d(e_1, e_3) + d(e_2, e_3)$.

As funções de distância Euclidiana e city block, derivadas da família Minkowski, são exemplos de distâncias métricas.

A teoria dos conjuntos difusos fornece uma base natural para a generalização das relações binárias entre dois objetos [Ovchinnikov 1991]. Relações de similaridade, um tipo particular de relações difusas, foram originalmente introduzidas por [Zadeh 1971] como uma generalização da definição clássica das relações de equivalência.

Uma relação binária difusa S no universo U é uma relação de similaridade em U se as seguintes propriedades forem satisfeitas [Zadeh 1971], [Ovchinnikov 1981] :

- Reflexividade: $S(x, x) = 1$, com $x \in U$;
- Simetria: $S(x, y) = S(y, x)$, com $x, y \in U$;
- Transitividade: $S(x, y) \geq S(x, z) * S(z, y)$, com $x, y, z \in U$ e $*$ uma T -norma.

Quando a transitividade não é requerida, ou seja, somente a reflexividade e a simetria são obedecidas, a relação difusa é chamada de relação de proximidade.

Na literatura pode-se encontrar diversos operadores para determinar o quanto dois conjuntos difusos são similares, como em [Zhang e Fu 2006], [Xuecheng 1992], [Chaudhuri e Rosenfeld 1999], [Wu e Mendel 2008], [Zwick et al. 1987], [Bustince et al. 2007], [Balopoulos et al. 2007] e [Wang 1997]. Todos estes operadores citados satisfazem apenas as propriedades de reflexividade e simetria,

Considerando o conjunto $X = \{x_1, x_2, \dots, x_n\}$ como o sendo o universo do discurso, e $CD(X)$ sendo o conjunto dos conjuntos difusos sobre este universo. Para dois conjuntos difusos $A, B \in CD(X)$, com as funções de pertinência μ_A e μ_B respectivamente, alguns operadores de similaridade são apresentados a seguir:

- Distância Euclideana normalizada [Zwick et al. 1987] é dada por:

$$DE(A, B) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \sqrt{(\mu_A(x_i) - \mu_B(x_i))^2} \quad (4.1)$$

- Distância city block normalizada [Zwick et al. 1987] é calculada conforme a seguir:

$$DC(A, B) = \frac{1}{n} \sum_{i=1}^n |\mu_A(x_i) - \mu_B(x_i)| \quad (4.2)$$

- Similaridade Cosseno [Wu e Mendel 2008] é obtida por:

$$SC(A, B) = \frac{\sum_{i=1}^n \mu_A(x_i) * \mu_B(x_i)}{(\sum_{i=1}^n \mu_A(x_i)^2)^{\frac{1}{2}} * (\sum_{i=1}^n \mu_B(x_i)^2)^{\frac{1}{2}}} \quad (4.3)$$

- Bustince et al. [Bustince et al. 2007] propuseram duas medidas para comparar conjuntos difusos que envolve a agregação dos operadores de implicação. Uma destas medidas denominada de Equality Index é dada por:

$$EQ_{DI}(A, B) = \text{Min} \{ \sigma_{DI}(A, B), \sigma_{DI}(B, A) \} \quad (4.4)$$

onde

$$\sigma_{DI}(A, B) = \frac{1}{n} \sum_{i=1}^n I(\mu_A(x_i), \mu_B(x_i)) \text{ e}$$

$$I(a, b) = \text{Min}(1, 1 - a + b) \quad - \text{ operador de implicação Lukasiewicz}$$

- Balopoulos et. al [Balopoulos et al. 2007] propuseram uma distância que utiliza a operação de agregação entre dois conjuntos difusos para gerar uma matriz, para depois aplicar uma norma. A medida de distância entre conjuntos difusos é calculada conforme a seguir:

$$DN(A, B; \sigma) = \|\Sigma(A, A) - \Sigma(B, B)\|_{n2} \quad (4.5)$$

onde $\Sigma(Z, Y)$ é uma matriz quadrada $n \times n$ sendo Z e Y dois conjuntos difusos de n elementos. O elemento i, j da matriz $\Sigma(Z, Y) = I(\mu_Z(x_i), \mu_Y(x_j))$ e

$$I(a, b) = \text{Min}(1, 1 - a + b) \quad - \text{ operador de implicação Lukasiewicz}$$

- Um operador de similaridade baseado na lógica difusa e nos operadores de implicação [Wu e Mendel 2008]:

$$\begin{aligned} [\mu_A(x_i) \iff \mu_B(x_i)] &\equiv [\mu_A(x_i) \longrightarrow \mu_B(x_i)] \wedge [\mu_B(x_i) \longrightarrow \mu_A(x_i)] \\ SL(A, B) &= \frac{1}{n} \sum_{i=1}^n [\mu_A(x_i) \iff \mu_B(x_i)] \end{aligned} \quad (4.6)$$

$$SL(A, B) = \frac{1}{n} \sum_{i=1}^n I[\mu_A(x_i), \mu_B(x_i)] \wedge I[\mu_B(x_i), \mu_A(x_i)]$$

onde $I(a, b) = \text{Min}(1, 1 - a + b)$ - operador de implicação Lukasiewicz

- Wang [Wang 1997] propôs a medida de similaridade entre conjuntos difusos mostrada a seguir:

$$SW(A, B) = \frac{1}{n} \sum_{i=1}^n [1 - |\mu_A(x_i) - \mu_B(x_i)|] \quad (4.7)$$

4.3 Considerações Finais do Capítulo

Neste capítulo foram apresentados alguns dos operadores de similaridades entre conjuntos difusos. Foram apresentados operadores de similaridades de várias características diferentes, como: operadores baseados em proximidades, operador baseado no cosseno entre os objetos, operador baseado na lógica difusa e outras características.

Os resultados produzidos por sistemas de classificação e de recuperação de imagens por conteúdo são fortemente dependentes da representação dos dados no espaço de característica e da medida de distância escolhida para determinar a similaridade entre os objetos do domínio da aplicação.

No próximo capítulo apresentamos uma nova proposta para representação de característica baseada em conjuntos difusos robusta quanto a escolha da medida de similaridade.

Capítulo 5

Representação do Espaço Visual de Característica Utilizando Conjuntos Difusos

5.1 Introdução

O caminho mais utilizado para a representação das características após a aplicação de um descriptor é a normalização. Neste capítulo é proposta uma nova representação para as características extraídas de uma imagem baseada em conjuntos difusos.

A representação de dados proposta tem como objetivo modelar a incerteza presente no processo de extração de característica.

5.2 Uma Nova Representação para o Espaço Visual de Característica

Considere Ix e Iy sendo a imagem de consulta e a imagem de alvo respectivamente. Seja $X = [x_1, x_2, \dots, x_n]$ e $Y = [y_1, y_2, \dots, y_n]$ os vetores de características para Ix e Iy , sendo x_i e y_i os valores numéricos que representam a i^{th} característica da imagem Ix e Iy respectivamente. Neste contexto, n representa o número de características visuais que representam uma imagem.

Na representação convencional das características, cada valor numérico x_i pertencente a X e y_i pertencente a Y , é normalizado para o intervalo $[0,1]$. Os vetores de características normalizados são agora representados como $X_p = [x_{p1}, x_{p2}, \dots, x_{pn}]$ e $Y_p = [y_{p1}, y_{p2}, \dots, y_{pn}]$. Uma medida de similaridade é aplicada aos vetores X_p e Y_p para calcular o quanto as imagens Ix e Iy são similares.

Apesar da simplicidade desta representação da imagem, as incertezas inerentes ao processo de extração de característica não são levadas em consideração. Visando uma

solução a este problema, neste trabalho é proposta representar cada característica em X e Y como um conjunto difuso, levando em consideração as funções de pertinência que compõem as partições difusas modeladas para cada característica. A representação proposta, passa a representar cada elemento de X e Y por conjuntos difusos, resultando em $X_f = [x_{f1}, x_{f2}, \dots, x_{fn}]$ e $Y_f = [y_{f1}, y_{f2}, \dots, y_{fn}]$. Note que os elementos x_{fi} e y_{fi} são conjuntos difusos que representam o i -ésimo elemento de X e Y respectivamente. A idéia da estratégia proposta é preservar a incerteza presente na fronteira entre dois conjuntos difusos adjacentes visando melhorar a busca por similaridade. O cálculo da similaridade entre as imagens Ix e Iy , considerando os vetores X_f e Y_f , é dado pela Definição 1.

Definição 1. A medida de similaridade S para os vetores X_f e Y_f é calculada pela média ponderada das distâncias entre cada par x_{fi} e y_{fi} .

$$S(X_f, Y_f) = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n (w_i * \text{sim}(x_{fi}, y_{fi})) \quad (5.1)$$

onde $\text{sim}(x_{fi}, y_{fi})$ é uma medida de similaridade aplicada a dois conjuntos difusos x_{fi} e y_{fi} . Os pesos w_i é associado com a característica i . O significado de S depende da métrica usada para calcular $\text{sim}(\cdot)$.

5.3 Um exemplo numérico

Para facilitar o entendimento da nova representação de dados, considere um espaço de característica bi-dimensional, ou seja $n = 2$, conforme mostrado na Figura 5.1 ('o' representa a classe 1 e 'x' representa a classe 2). Note que o espaço característica é bem clusterizável para a característica 2 (d_2), o mesmo não ocorre na característica 1 (d_1).

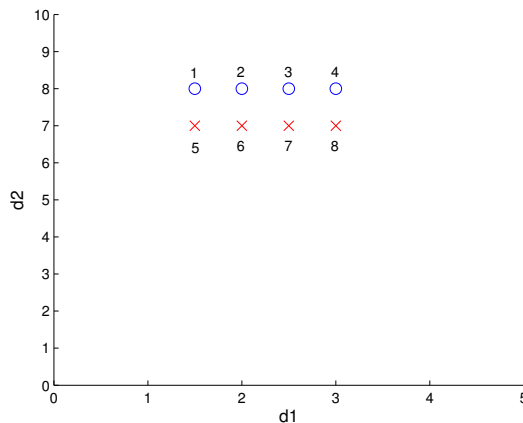


Figura 5.1: Um exemplo numérico simples. O domínio de d_1 é o intervalo $[0,5]$ e o domínio de d_2 é o intervalo $[0,10]$.

Considerando o_1 como o objeto consulta, foi aplicado duas abordagens de representação para calcular a similaridade entre o objeto consulta e os outros objetos da base de

dados (Figura 5.1):

- **representando as características pelos valores normalizados** - cada característica é normalizada no intervalo $[0,1]$ considerando o domínio $[0, 5]$ para a característica d1 e o domínio $[0, 10]$ para a característica d2. A representação normalizada é mostrada na Tabela 5.1. Foram aplicadas três medidas de similaridades (distância Euclideana, distância city block e distância cosseno) e o resultado da busca é apresentado na ordem crescente de similaridade na Tabela 5.2. Note que a ordem dos elementos retornados e consequentemente a precisão da resposta varia de acordo com a escolha da medida de similaridade.
- **representando as características por conjuntos difusos** - considere a partição difusa mostrada na Figura 5.2. A representação proposta obtida a partir destas partições é exibida na Tabela 5.1. A distância Euclidiana foi escolhida como a medida de similaridade para calcular $sim(\cdot)$, como definido na Equação 5.1. Os pesos atribuídos a cada partição (w_i) são indicados no título da Figura 5.2. Maiores detalhes sobre a definição dos valores dos pesos serão apresentados no capítulo 6. O resultado é apresentado na Tabela 5.2. Note que a representação proposta consegue melhorar a separabilidade entre os exemplares das diferentes classes, melhorando assim os resultados da busca por similaridade (a precisão na busca por similaridade foi de 100%). Aplicando a distância city block e a distância cosseno, os resultados no ranking são exatamente os mesmo quando se utiliza a representação das características por conjuntos difusos proposta neste trabalho, diferente do que ocorreu quando a representação normalizada foi usada, veja a Tabela 5.2.

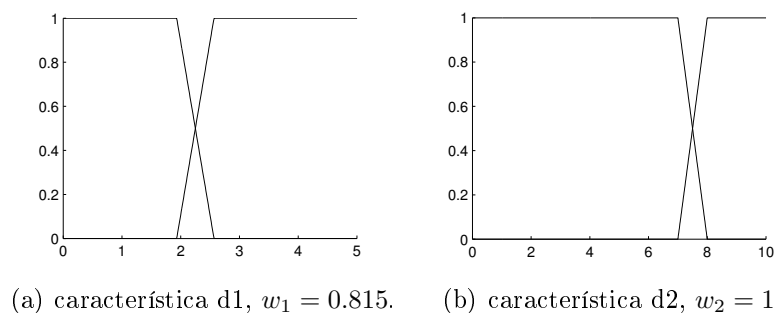


Figura 5.2: Partição difusa para cada dimensão do espaço característica mostrado na Figura 5.1.

Tabela 5.1: Diferentes representações para o espaço característica da Figura 5.1.

objetos	espaço original		espaço normalizado		repres. conjuntos difusos	
	d_1	d_2	d_1	d_2	d_1	d_2
o_1	1.5	8	0.3	0.8	$\{1,0\}$	$\{0,1\}$
o_2	2.0	8	0.4	0.8	$\{0.89,0.10\}$	$\{0,1\}$
o_3	2.5	8	0.5	0.8	$\{0.10,0.89\}$	$\{0,1\}$
o_4	3.0	8	0.6	0.8	$\{0,1\}$	$\{0,1\}$
o_5	1.5	7	0.3	0.7	$\{1,0\}$	$\{1,0\}$
o_6	2.0	7	0.4	0.7	$\{0.89,0.10\}$	$\{1,0\}$
o_7	2.5	7	0.5	0.7	$\{0.10,0.89\}$	$\{1,0\}$
o_8	3.0	7	0.6	0.7	$\{0,1\}$	$\{1,0\}$

Tabela 5.2: Resultados da busca por similaridades obtidos para as diferentes representações do espaço, considerando o_1 como o objeto consulta.

	ranking dos resultados							
<i>distância Euclideana</i>								
repres. normalizada	o_2	o_5	o_6	o_3	o_7	o_4	o_8	
repres. conjuntos difusos	o_2	o_3	o_4	o_5	o_6	o_7	o_8	
<i>distância city block</i>								
repres. normalizada	o_2	o_5	o_3	o_6	o_4	o_7	o_8	
repres. conjuntos difusos	o_2	o_3	o_4	o_5	o_6	o_7	o_8	
<i>distância coseno</i>								
repres. normalizada	o_5	o_2	o_6	o_3	o_7	o_4	o_8	
repres. conjuntos difusos	o_2	o_3	o_4	o_5	o_6	o_7	o_8	

5.4 Considerações Finais do Capítulo

Neste capítulo foi proposta uma nova representação do espaço característica. A nova representação trata cada característica como um conjunto difuso, visando modelar a incerteza que existe nos processos de extração de características, o que não ocorre na representação por valores normalizados. Neste capítulo foi apresentada uma estratégia para obter a medida de similaridade entre objetos cujas características são representadas de acordo com a representação proposta.

No exemplo numérico apresentado neste capítulo, uma das dimensões é perfeitamente agrupável enquanto a outra apresenta pouca discriminação entre os objetos. Mesmo após a aplicação de métodos para redução do espaço de características, situação como esta pode se manter. A representação proposta é capaz de produzir bons resultados na busca por similaridade, mesmo sob condições indesejáveis, como ilustrado no exemplo. Avaliação com bases de dados reais, a serem apresentadas no capítulo 7, confirmam a robustez da representação de dados proposta.

É importante destacar que a nova representação é fortemente dependente da partição utilizada e que devido a isto nós também propomos um método automático e não supervisionado para obtenção de partição difusa a partir de uma base de dados, a ser descrito no próximo capítulo.

Capítulo 6

Método Automático e Não-Supervisionado Para a Geração de Partições Difusas (MANGPD)

6.1 Introdução

A geração automática de partições difusas a partir da base de dados apresenta-se como uma alternativa à partições difusas definidas por um especialista. Quando não há disponibilidade de um especialista para um determinado domínio, a definição das funções de pertinência só é possível via técnicas de aprendizagem de máquina.

A principal idéia do método para geração automática de partições difusas é utilizar-se da distribuição dos dados para a definição da partição. Utilizando-se do algoritmo de agrupamento de dados Fuzzy C-means pode-se definir as partições com funções de pertinência tendo formas triangulares, trapezoidais ou semi-trapezoidais.

Alguns dos métodos descritos no Capítulo 3 utilizam de técnicas supervisionadas, outros se limitam a gerar funções de pertinência apenas triangulares com alta sobreposição entre estas funções. No nosso método, as formas das funções de pertinência não está limitada apenas a triangular, as funções de pertinência não possuem alta sobreposição entre si e também não utiliza-se de técnicas supervisionadas.

6.2 Método Proposto

O método proposto para gerar as partições difusas é baseado no algoritmo de agrupamento Fuzzy C-Means (FCM). Para cada dimensão do espaço de característica é encontrado sua respectiva partição, sendo cada característica tratada de forma independente.

No método proposto, o algoritmo FCM é aplicado em cada eixo do espaço de característica. A Figura 6.1 ilustra um possível resultado do FCM em uma dimensão do espaço

de características. Note que um mesmo valor pertence a diferentes agrupamentos, com diferentes graus de pertinência.

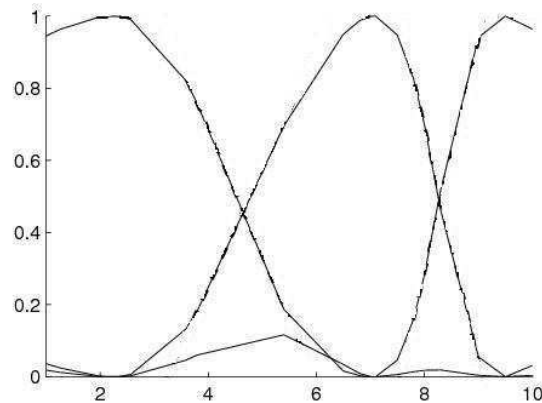


Figura 6.1: Resultado do FCM em uma dimensão do espaço de característica, no qual foram encontrados três agrupamentos. O valor 6 possui os graus de pertinência 0.18 em relação ao primeiro agrupamento, 0.72 em relação ao segundo agrupamento e 0.10 em relação ao terceiro agrupamento.

A partir do resultado final de um processo de agrupamento, a partição difusa é obtida. Considerando dois agrupamentos adjacentes obtidos, é necessário identificar o ponto de intersecção entre os mesmos. Para identificar o ponto de intersecção, deve-se encontrar as equações das duas retas de primeiro grau, sendo uma com coeficiente angular positivo e a outra com coeficiente angular negativo.

A determinação das equações das retas é feita a partir de dois pontos conhecidos das mesmas. Para a primeira reta, os pontos conhecidos são $(x_1, \mu_2(x_1))$ e $(x_2, \mu_2(x_2))$ e para a segunda reta os pontos conhecidos são $(x_1, \mu_1(x_1))$ e $(x_2, \mu_1(x_2))$, sendo x_1 e x_2 os valores do eixo, $\mu_2(x_1)$ o valor de pertinência de x_1 ao grupo 2, $\mu_2(x_2)$ o valor de pertinência de x_2 ao grupo 2, $\mu_1(x_1)$ o valor de pertinência de x_1 ao grupo 1 e $\mu_1(x_2)$ o valor de pertinência de x_2 ao grupo 1. A identificação dos pontos das retas é mostrada na Figura 6.2

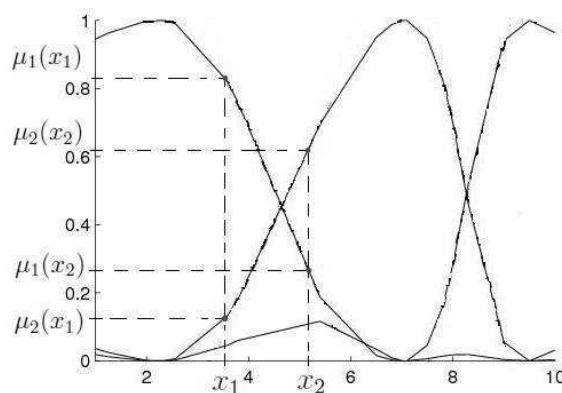


Figura 6.2: Pontos para identificação das retas para a geração de duas funções de pertinência adjacentes, tendo como base um resultado do FCM.

Uma vez identificadas, as retas são alongadas nas duas direções: até alcançarem o eixo X e até o ponto $\mu(x_i) = 1$. A forma das funções de pertinência depende do ponto de

intersecção entre os grupos adjacentes, podendo ter formas triangulares ou trapezoidal. Nas funções das extremidades, as funções de pertinência são semi trapezoidal.

Note que as funções de pertinência podem não ser simétricas e que as mesmas possuem uma área de sobreposição entre os conjuntos difusos menor do que os métodos descritos no capítulo 3. Também, é importante destacar que os parâmetros das funções de pertinência não são pré-definidos como em outros trabalhos relacionados.

Aplicando o método proposto sobre o resultado do Fuzzy C-Means ilustrado na Figura 6.1, uma possível partição é ilustrada na Figura 6.3.

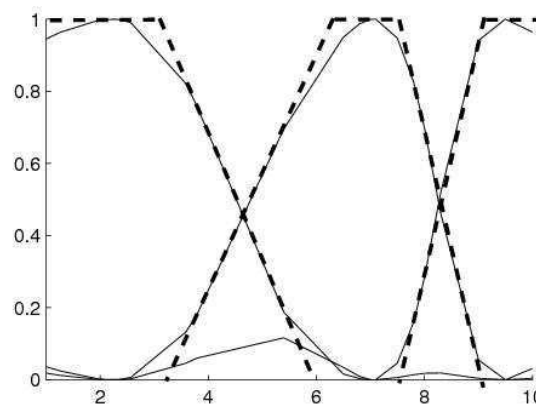


Figura 6.3: Partição difusa obtida pelo método proposto. A linha sólida indica o resultado do algoritmo FCM e a linha tracejada indica a funções de pertinência encontradas pelo método proposto.

Uma das características do algoritmo FCM é a necessidade de definir previamente o número K de agrupamentos desejados. Com o objetivo de encontrar a melhor partição difusa para uma dada dimensão, o algoritmo FCM é executado para K variando de dois a sete (mesma estratégia usada em [Guillaume e Charnomordic 2004]). A ideia de limitar K em sete é garantir a interpretabilidade do resultado e reduzir o tempo computacional.

Tendo vários agrupamentos possíveis para uma dada dimensão, surge a seguinte pergunta: - Como escolher o melhor agrupamento? Para responder a esta questão, nós propomos a utilização de uma medida de validação de agrupamento não supervisionada, conhecida como coeficiente de silhueta. Conforme visto na Seção 2.4.2, este coeficiente expressa a relação de medidas de coesão e separação dentro dos grupos e também entre os grupos. A melhor partição é obtida a partir do agrupamento que possuir o maior coeficiente de silhueta.

O coeficiente de silhueta do melhor agrupamento resultante será utilizado como peso (w_i em Equação 5.1) para a respectiva partição.

O algoritmo para o método proposto é descrito a seguir:

Algorithm 3 Algoritmo do método proposto para geração de partições difusas

```

1: for all dimensão  $d_i$ ,  $i = 1$  até  $N$  do
2:   for all  $NroCluster = 2$  até  $7$  do
3:      $FCM(d_i, NroCluster)$ 
4:     Armazena o resultado
5:     Calcula o coeficiente de silhueta  $csNrocluster$ 
6:     Armazena o  $csNroCluster$ 
7:   end for
8:   Define o  $NroCluster$  do melhor agrupamento // De acordo com o  $csNroCluster$ 
    máximo
9:   Define o resultado do melhor agrupamento
10:  Encontra os pontos de intersecção entre os cluster adjacentes
11:  Determina as equações das retas
12:  Determina os parâmetros das funções de pertinência difusas e as armazena
13:  Define a partição para a dimensão  $d_i$ 
14: end for

```

6.3 Considerações Finais do Capítulo

Para obter a representação de dados baseada em conjuntos difusos conforme proposta, é necessário que se defina as partições difusas para cada dimensão do espaço característica. Neste capítulo foi apresentado um método para obtenção das partições difusas a partir da base de dados. O método apresentado é automático e não-supervisionado e utiliza-se da distribuição dos dados em cada eixo através do uso do algoritmo FCM, as formas das funções de pertinência não estão limitadas apenas a triangular e as funções de pertinência não possuem alta sobreposição entre si.

Para a validação das propostas deste trabalho foram realizados alguns experimentos empregando os conceitos apresentados e os resultados são mostrados no próximo capítulo.

Capítulo 7

Avaliação e Discussão dos Resultados

7.1 Introdução

Para a avaliação da representação de dados proposta e também do método proposto para a geração automática de partições difusas foram realizados experimentos com bases de dados sintéticos e reais. Os experimentos foram realizados visando comparar o comportamento da representação normalizada com a representação proposta perante diferentes operadores de similaridades, e comparar o desempenho do método proposto para a geração das partições difusas com outros métodos similares.

7.2 Metodologia utilizada nos experimentos

Para avaliar a qualidade da recuperação dos dados foram utilizadas as curvas de Precisão/Revocação e também a Mean Average Precision (MAP) [Manning et al. 2008].

Em todos os experimentos foi utilizada a estratégia de validação *leave-one-out*.

Os experimentos realizados possuem dois objetivos principais:

- **comparar a representação proposta com a representação normalizada** - para cada base de dados foi obtida a representação normalizada e também a representação das características por conjuntos difusos (definida através do método proposto). Foram utilizados os sete operadores de similaridades mostrados na Seção 4.2: Distância Euclidiana (DE), Distância City block (DC), Similaridade cosseno (SC), Equality index (EQ), Distância baseado em norma (DN), Similaridade baseado em lógica (SL) e Similaridade proposta por Wang (SW);
- **comparar o método proposto para geração das partições difusas com métodos similares** - em cada base de dados foi obtida a representação proposta utilizando os métodos Partição Difusa Uniforme (PDU), Partição Difusa utilizando o K-means (PDK), Hierarchical Fuzzy Partitioning (HFP) e o método proposto (Mé-

todo Automático e Não-Supervisionado Para a Geração de Partição Difusa descrito no capítulo 6). Para esta comparação foi utilizado a Distância Euclidiana.

Nas Seções 7.3 e 7.4 são destacadas as bases de dados utilizadas, os resultados obtidos e a discussão dos mesmos.

7.3 Base de Dados Sintéticas

Foram utilizadas três bases de dados sintéticas, que são tratadas aqui como BDSintetica01, BDSintetica02 e BDSintetica03. Para a geração das bases de dados sintéticas foi utilizado o programa DBGen [DBGen].

A base de dados BDSintetica01 é composta de 40 elementos, classificados em duas classes com 20 elementos cada, sendo os atributos bem clusterizáveis. A figura 7.1 mostra esta base de dados.

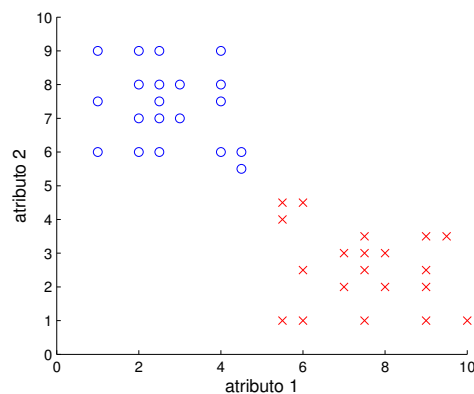


Figura 7.1: Base de dados BDSintetica01.

A base de dados BDSintetica02 é composta de 150 elementos, classificados em três classes com 50 elementos cada, sendo os elementos bem clusterizáveis em relação ao atributo 1. A figura 7.2 mostra esta base de dados.

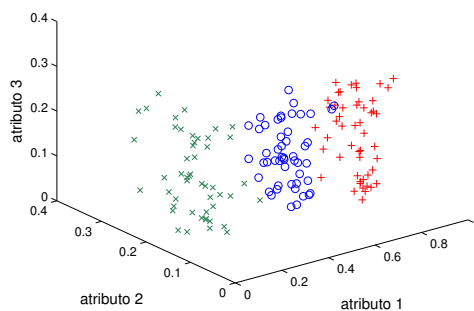


Figura 7.2: Base de dados BDSintetica02.

A base de dados BDSintetica03 é composta de 200 elementos, classificados em quatro classes com 50 elementos cada, sendo os elementos bem clusterizáveis. A figura 7.3 mostra esta base de dados.

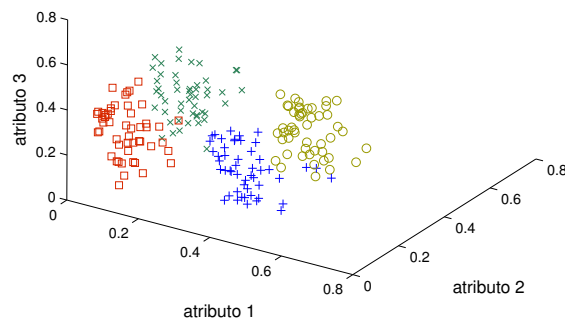
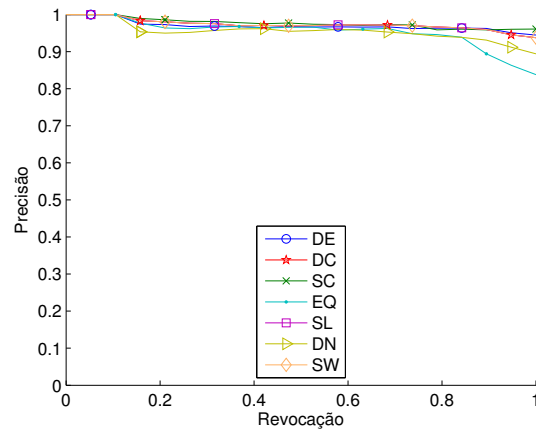
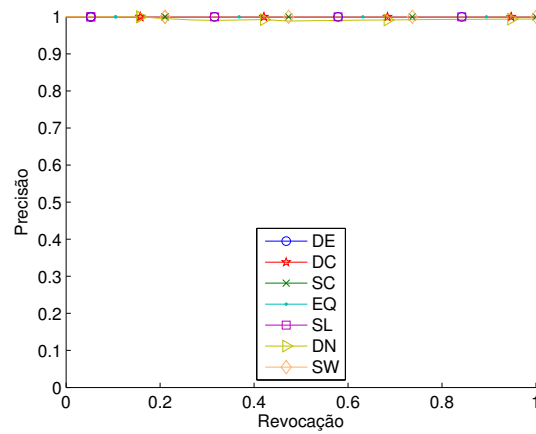


Figura 7.3: Base de dados BDSintetica03.

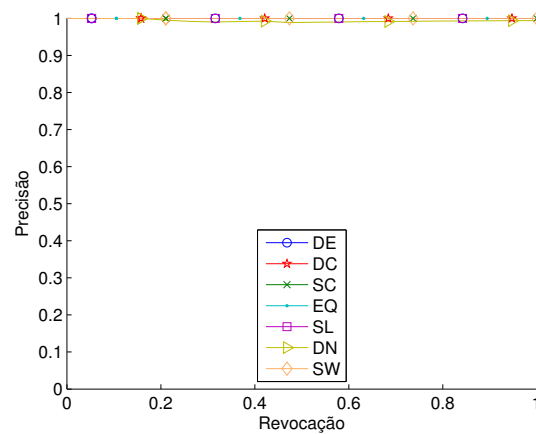
As Figuras 7.4, 7.5 e 7.6 mostram as curvas de precisão/revocação para os experimentos de comparação entre as representações das características utilizando vários operadores de similaridades e as bases de dados BDSintetica01, BDSintetica02 e BDSintetica03 respectivamente. As Mean Average Precision (MAP) relativas a estes experimentos são mostradas na Tabela 7.1. Os resultados mostram que utilizando as propostas deste trabalho obteve-se os melhores resultados em todos os experimentos. Para a base de dados BDSintetica01 em particular, utilizando a representação proposta, foi possível obter o resultado ideal para a curva de precisão/revocação para seis operadores de similaridades. Outro fato importante é que a representação das características por conjuntos difusos mostrou-se menos sensível à escolha da medida de similaridade, que é um requisito importante para sistemas CBIR.



(a) Experimento 1: representação normalizada.

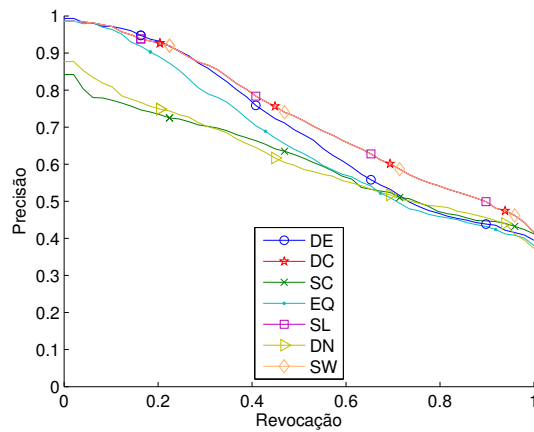


(b) Experimento 2: representação das características utilizando conjuntos difusos. Medidas de similaridades sem ponderação, $w_i = 1$.

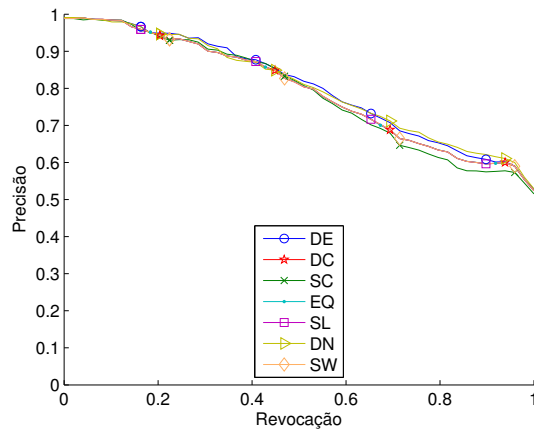


(c) Experimento 3: representação das características utilizando conjuntos difusos. Medidas de similaridades com ponderação, $w_1 = 0.8185$, $w_2 = 0.8595$.

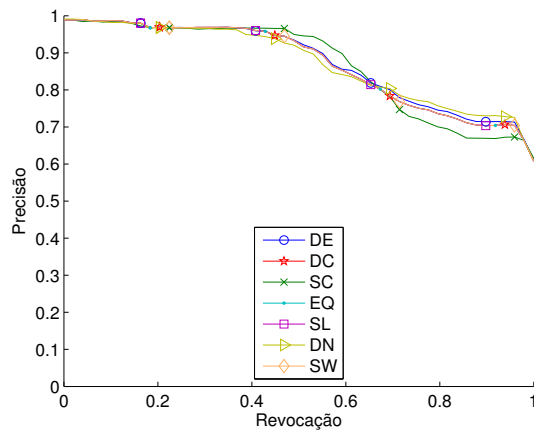
Figura 7.4: Experimentos usando a base de dados sintética BDSintetica01.



(a) Experimento 1: representação normalizada.

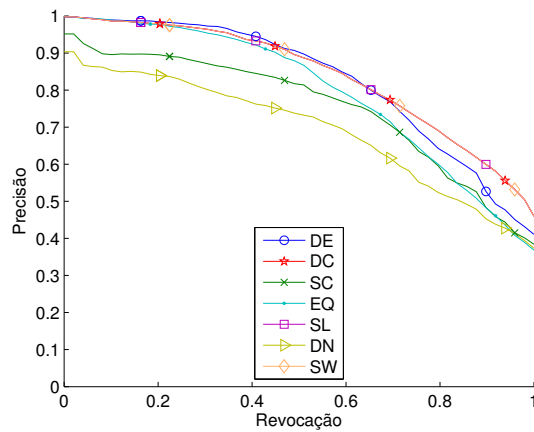


(b) Experimento 2: representação das características utilizando conjuntos difusos. Medidas de similaridades sem ponderação, $w_i = 1$.

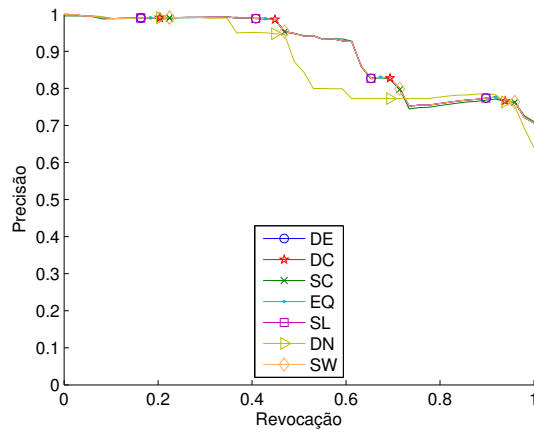


(c) Experimento 3: representação das características utilizando conjuntos difusos. Medidas de similaridades com ponderação, $w_1 = 0.9149$, $w_2 = 0.7228$, $w_3 = 0.7911$.

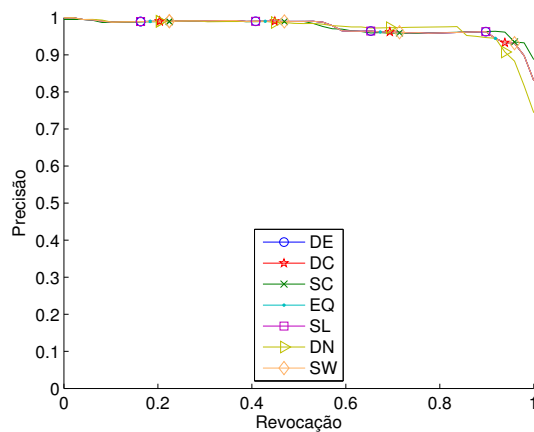
Figura 7.5: Experimentos usando a base de dados sintética BDSintetica02.



(a) Experimento 1: representação normalizada.



(b) Experimento 2: representação das características utilizando conjuntos difusos. Medidas de similaridades sem ponderação, $w_i = 1$.



(c) Experimento 3: representação das características utilizando conjuntos difusos. Medidas de similaridades com ponderação, $w_1 = 0.93341$, $w_2 = 0.94693$, $w_3 = 0.79224$.

Figura 7.6: Experimentos usando a base de dados sintética BDSintetica03.

Tabela 7.1: Mean Average Precision dos experimentos: (1) representação normalizada. (2) representação das características utilizando conjuntos difusos sem ponderação no cálculo da similaridade. (3) representação das características utilizando conjuntos difusos com ponderação no cálculo da similaridade. Também é mostrado o ganho relativo da utilização da representação proposta em relação a representação normalizada.

	experimento 1	experimento 2	experimento 3	ganho relativo
<i>BDSintetica01</i>				
Distância Euclideana	96.9107	100	100	3.18 %
Distância City block	97.2464	100	100	2.83 %
Similaridade Cosseno	97.6053	100	100	2.45 %
Equality Index	95.3343	100	100	4.89 %
Distância baseada em norma	95.3328	99.3133	99.3133	4.17 %
Similaridade baseado em lógica	97.2464	100	100	2.83 %
Similaridade proposta por Wang	97.2464	100	100	2.83 %
<i>BDSintetica02</i>				
Distância Euclideana	68.8748	80.1427	87.1405	26.52 %
Distância City block	72.2608	79.0578	86.7339	20.02 %
Similaridade Cosseno	60.8225	78.4269	86.4259	42.09 %
Equality Index	65.9337	79.0602	86.7339	31.54 %
Distância baseada em norma	60.9376	80.1302	86.947	42.68 %
Similaridade baseado em lógica	72.2608	79.0626	86.7339	20.02 %
Similaridade proposta por Wang	72.2608	79.0626	86.7339	20.02 %
<i>BDSintetica03</i>				
Distância Euclideana	82.2229	89.8739	97.2612	18.29 %
Distância City block	83.4187	89.8739	97.2612	16.59 %
Similaridade Cosseno	74.834	79.29	97.4328	30.19 %
Equality Index	79.1022	89.8739	97.2612	22.95 %
Distância baseada em norma	68.6901	86.9674	96.9753	41.17 %
Similaridade baseado em lógica	83.4187	89.8739	97.2612	16.59 %
Similaridade proposta por Wang	83.4187	89.8739	97.2612	16.59 %

Neste ponto podemos avaliar o comportamento do método proposto para geração da partição difusa com relação aos outros métodos detalhados na Seção 3.3. Para todas as abordagens, a melhor partição é obtida utilizando o coeficiente de silhueta. O número de partições obtidas em cada dimensão e também a entropia dos agrupamentos resultantes são mostrados na Tabela 7.2 e o MAP computado nos experimentos é mostrado na Tabela 7.3.

A distribuição dos dados nas bases de dados sintéticas propiciam uma boa separabilidade entre as classes através da aplicação de técnicas de agrupamentos, o que pode ser verificado pelos valores de entropia mostrados na Tabela 7.2. Neste cenário, o método proposto para geração de partição difusa obteve os melhores resultados em todos os casos, podendo ser dado um destaque ao fato de que foi o único método que propiciou a MAP ideal para a base de dados BDSintetica01.

Tabela 7.2: Melhores partições difusas para cada característica.

Métodos	# Características visuais			Entropia
	1	2	3	
<i>Base de dados BDSintetica01</i>				
MANGPD	2	2	-	0
PDU	5	2	-	0
PDK	2	2	-	0
HFP	2	2	-	0.1208
<i>Base de dados BDSintetica02</i>				
MANGPD	3	3	3	0.6639
PDU	3	2	3	0.6494
PDK	3	3	2	0.6614
HFP	2	2	2	0.8274
<i>Base de dados BDSintetica03</i>				
MANGPD	2	2	2	0.6815
PDU	2	2	2	0.6659
PDK	2	2	2	0.6808
HFP	2	2	2	0.6817

Tabela 7.3: Mean Average Precision utilizando diferentes partições difusas.

base de dados	MANGPD	PDU	PDK	HFP
BDSintetica01	100	98.0118	98.6682	97.7311
BDSintetica02	87.1405	83.0468	83.8589	75.1207
BDSintetica03	97.2612	90.3585	91.2453	90.1757

7.4 Base de Dados Reais

Foram utilizadas cinco bases de dados reais: base de dados IRIS [UCI], base de dados WDBC [UCI], uma de base de dados sobre casos de câncer de mama referenciada neste trabalho como BD111, base de dados BDFACES [Nordstrøm et al. 2004] e base de dados BDSHAPES [Shapes].

A base de dados IRIS é bem utilizada em reconhecimento de padrões e contém informações sobre uma flor. É composta de 150 elementos classificados em 3 classes, sendo cada elemento representado por um vetor de característica de tamanho igual a 4.

A base de dados WDBC contém informações sobre casos de análise de células do nódulo de câncer de mama. É composta de 569 elementos classificados em 2 classes (casos benignos e casos malignos). As características que compõe o vetor de característica foram obtidas a partir de imagens digitalizadas de células do nódulo obtidas a partir da técnica de aspiração por agulha fina, descrevendo basicamente características dos núcleos das células presentes na imagem.

A base BD111 é uma base de dados obtidas a partir da digitalização de regiões de interesse do mamograma. Do total de imagens, 57 foram obtidas a partir do Screen Test: Programa de Detecção de câncer de mama de Alberta no Canadá [ScreenTest] e 54 foram obtidas da base de dados da Sociedade de Análise de Imagens Mamográficas do Reino Unido [MIAS]. No total esta base é composta de 111 imagens, sendo 65 lesões benignas e 46 tumores malignos.

O vetor de característica foi composto por descritores de formas obtidos a partir da função tangencial suavizada da aproximação poligonal de um contorno. As características presentes no vetor de características são: índice de espículos, duas diferentes medidas da dimensão fractal, índice de presença de regiões côncavas, índice de presença de regiões convexas e índice de convexidade. A forma de extração das característica é descrita no trabalho [Guliatto et al. 2008].

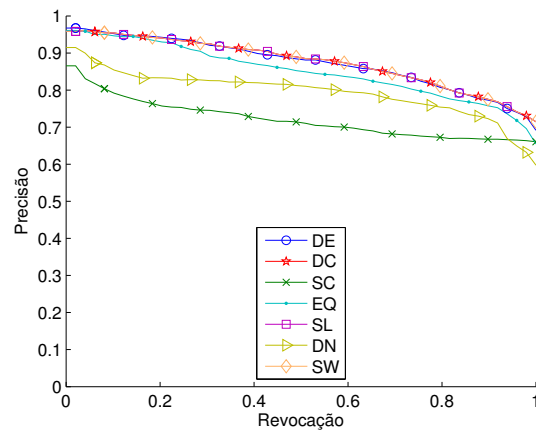
A base de dados BDFACES é um conjunto com 240 imagens de faces, utilizadas para reconhecimento digital de faces, divididas em 40 classes. As características extraídas foram 20 descritores de Fourier do histograma da imagem, 4 descritores de Fourier da imagem como um todo, média de intensidade da imagem e desvio padrão das intensidades da imagem.

A base de dados BDSHAPES é um conjunto com 340 imagens binárias contendo formas de objetos, divididas em 17 classes. Assim como na base de dados BDFACES, as características extraídas foram 20 descritores de Fourier do histograma da imagem, 4 descritores de Fourier da imagem como um todo, média de intensidade da imagem e desvio padrão das intensidades da imagem.

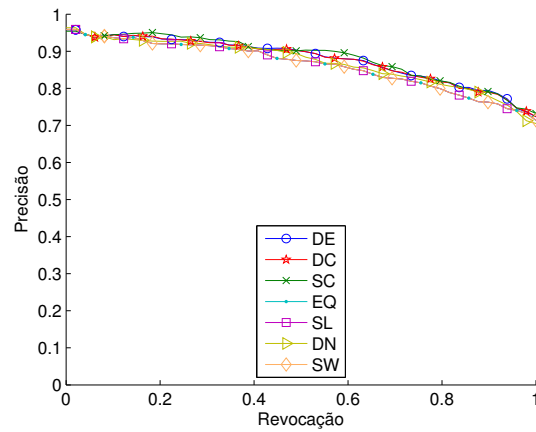
As Figuras 7.7, 7.8, 7.9, 7.10 e 7.11 mostram as curvas de precisão/revocação para os experimentos referente a comparação entre as representações das características utilizando vários operadores de similaridades e as bases de dados IRIS, WDBC, BD111, BDFACES e BDSHAPES. As Mean Average Precision (MAP) relativa a estes experimentos são mostradas na Tabela 7.4. Através da análise do resultados, pode-se perceber que utilizando a representação dos dados através de conjuntos difusos obteve-se os melhores resultados nas bases de dados IRIS, WDBC e BD111.

Na Tabela 7.5 são mostrados o número de funções de pertinência difusas para as melhores partições em cada característica e também é mostrada a entropia dos agrupamentos. A representação dos dados proposta é fortemente dependente do método de geração de partições difusas, o qual está baseado na técnica de agrupamento Fuzzy C-Means. Para as bases de dados BDFACES e BDSHAPES, não foi possível através do Fuzzy C-Means obter um agrupamento resultante em que os agrupamentos fossem bem homogêneos em relação à classe dos objetos, o que pode ser notado pelo valor da entropia. Nestas bases de dados, em que o agrupamento resultante possui uma entropia com um valor alto, a representação proposta não conseguiu obter melhor resultado em relação à representação normalizada.

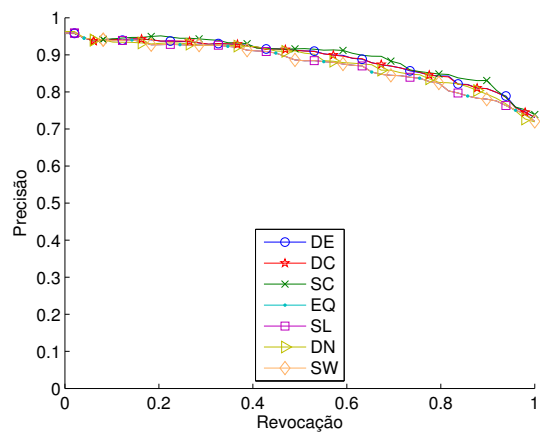
Também foi possível observar que a representação das características proposta neste trabalho mostrou-se menos sensível à escolha da medida de similaridade em todas as bases de dados, conforme já havia sido apresentado nos experimentos realizados com as bases de dados sintéticas.



(a) Experimento 1: representação normalizada.

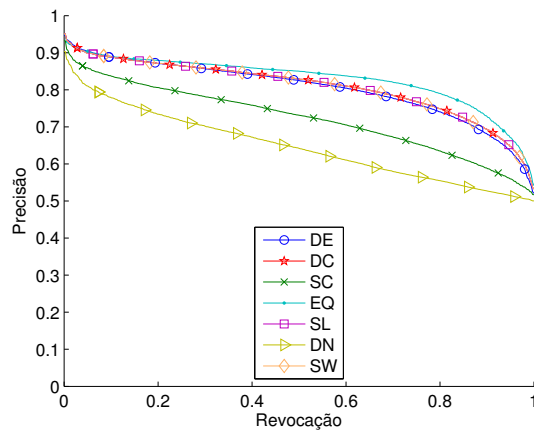


(b) Experimento 2: representação das características utilizando conjuntos difusos. Medidas de similaridades sem ponderação, $w_i = 1$.

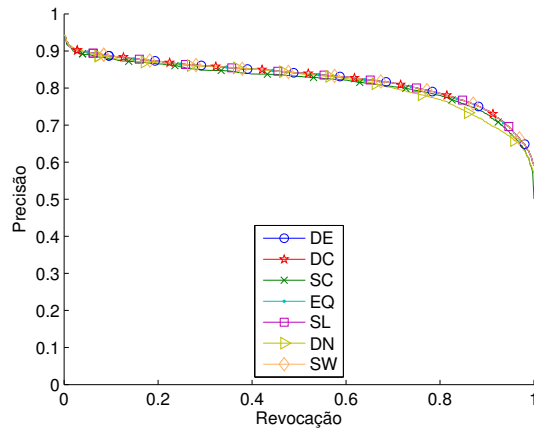


(c) Experimento 3: representação das características utilizando conjuntos difusos. Medidas de similaridades com ponderação, $w_1 = 0.7574$, $w_2 = 0.7137$, $w_3 = 0.9010$, $w_4 = 0.8584$.

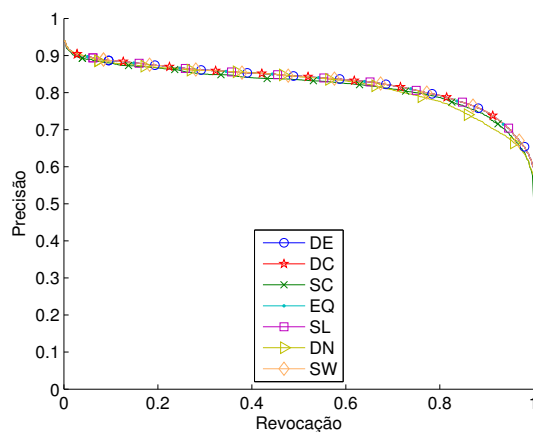
Figura 7.7: Experimentos usando a base de dados IRIS.



(a) Experimento 1: representação normalizada.

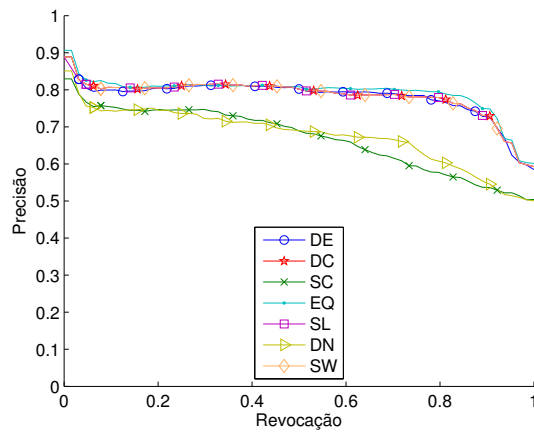


(b) Experimento 2: representação das características utilizando conjuntos difusos. Medidas de similaridades sem ponderação, $w_i = 1$.

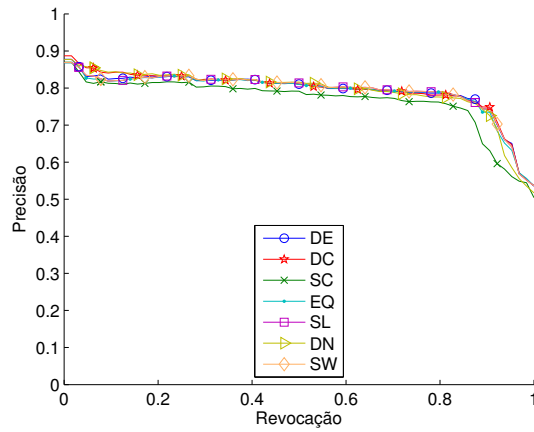


(c) Experimento 3: representação das características utilizando conjuntos difusos. Medidas de similaridades com ponderação, $w_1 = 0.6872$, $w_2 = 0.7265$, $w_3 = 0.8034$, $w_4 = 0.7114$, $w_5 = 0.7993$, $w_6 = 0.8456$, $w_7 = 0.7065$, $w_8 = 0.7455$, $w_9 = 0.8025$, $w_{10} = 0.8014$.

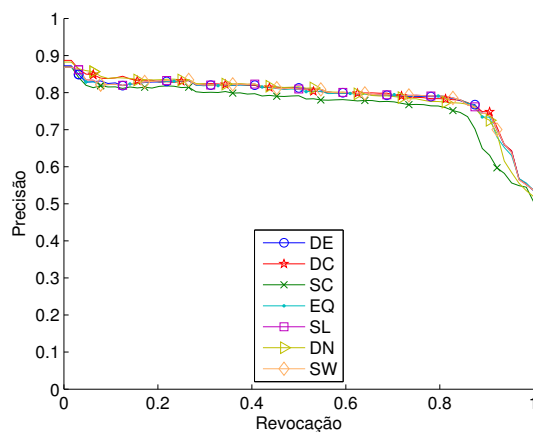
Figura 7.8: Experimentos usando a base de dados WDBC.



(a) Experimento 1: representação normalizada.

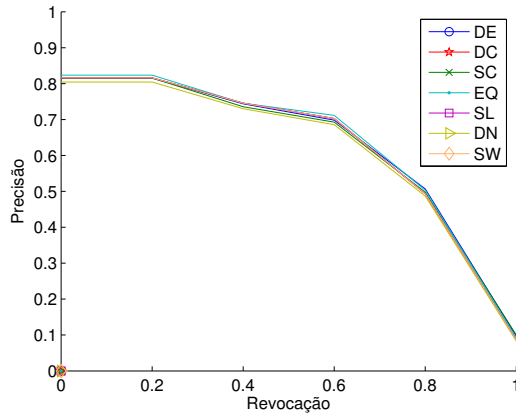


(b) Experimento 2: representação das características utilizando conjuntos difusos. Medidas de similaridades sem ponderação, $w_i = 1$.

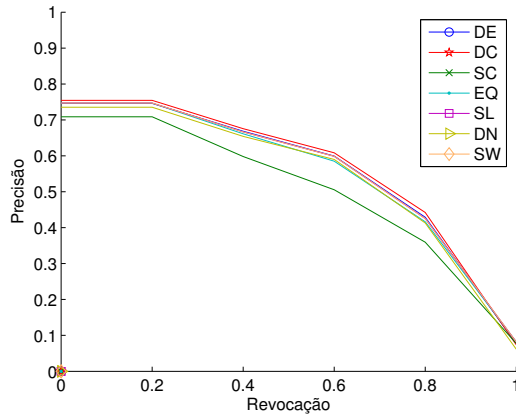


(c) Experimento 3: representação das características utilizando conjuntos difusos. Medidas de similaridades com ponderação, $w_1 = 0.87959$, $w_2 = 0.97984$, $w_3 = 0.83411$, $w_4 = 0.84202$, $w_5 = 0.88024$, $w_6 = 0.79375$, $w_7 = 0.80162$, $w_8 = 0.79415$.

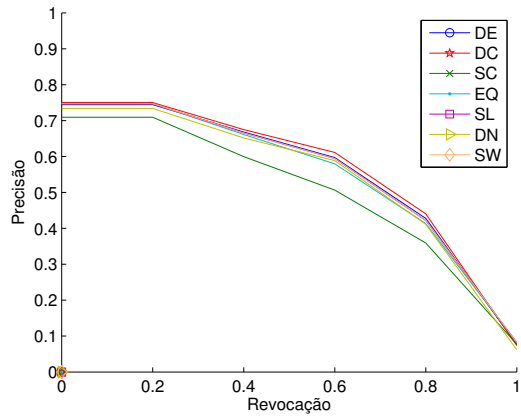
Figura 7.9: Experimentos usando a base de dados BD111.



(a) Experimento 1: representação normalizada.

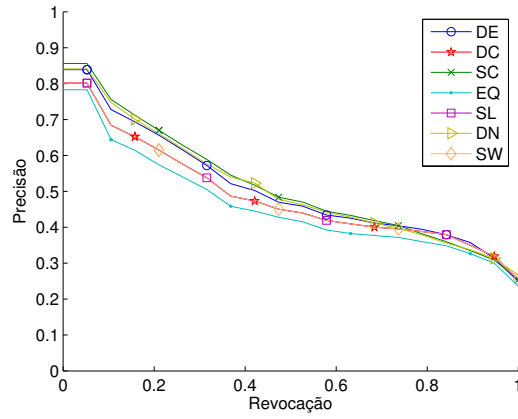


(b) Experimento 2: representação das características utilizando conjuntos difusos.
Medidas de similaridades sem ponderação, $w_i = 1$.

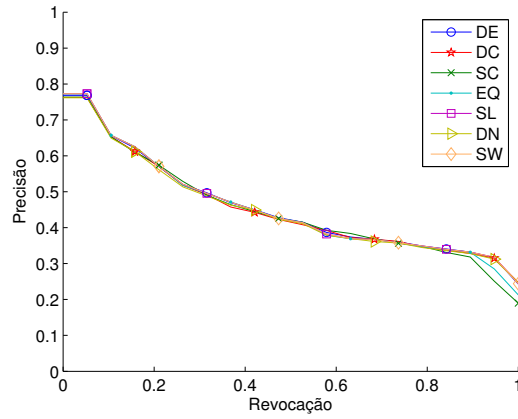


(c) Experimento 3: representação das características utilizando conjuntos difusos.
Medidas de similaridades com ponderação,
 $w_1 = 0.77515$, $w_2 = 0.79022$, $w_3 = 0.81359$,
 $w_4 = 0.83124$, $w_5 = 0.76204$, $w_6 = 0.85734$,
 $w_7 = 0.76374$, $w_8 = 0.78729$, $w_9 = 0.82041$,
 $w_{10} = 0.76176$, $w_{11} = 0.76176$, $w_{12} = 0.82041$,
 $w_{13} = 0.78729$, $w_{14} = 0.76374$, $w_{15} = 0.85734$,
 $w_{16} = 0.76204$, $w_{17} = 0.83124$, $w_{18} = 0.81359$,
 $w_{19} = 0.79022$, $w_{20} = 0.77515$, $w_{21} = 0.73643$,
 $w_{22} = 0.75318$, $w_{23} = 0.75393$, $w_{24} = 0.75989$,
 $w_{25} = 0.75919$, $w_{26} = 0.79432$, $w_{27} = 0.76148$,
 $w_{28} = 0.76441$.

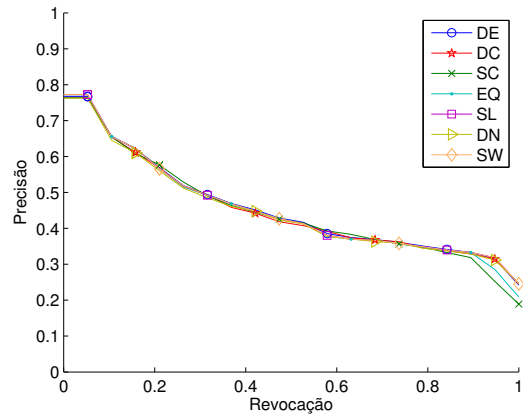
Figura 7.10: Experimentos usando a base de dados BDFACES.



(a) Experimento 1: representação normalizada.



(b) Experimento 2: representação das características utilizando conjuntos difusos.
Medidas de similaridades sem ponderação, $w_i = 1$.



(c) Experimento 3: representação das características utilizando conjuntos difusos.
Medidas de similaridades com ponderação,
 $w_1 = 0.84343$, $w_2 = 0.84208$, $w_3 = 0.84171$,
 $w_4 = 0.84098$, $w_5 = 0.83896$, $w_6 = 0.83538$,
 $w_7 = 0.83463$, $w_8 = 0.83212$, $w_9 = 0.83078$,
 $w_{10} = 0.82528$, $w_{11} = 0.82528$, $w_{12} = 0.83078$,
 $w_{13} = 0.83212$, $w_{14} = 0.83463$, $w_{15} = 0.83538$,
 $w_{16} = 0.83896$, $w_{17} = 0.84098$, $w_{18} = 0.84171$,
 $w_{19} = 0.84208$, $w_{20} = 0.84343$, $w_{21} = 0.75705$,
 $w_{22} = 0.73607$, $w_{23} = 0.73151$, $w_{24} = 0.72567$,
 $w_{25} = 0.78033$, $w_{26} = 0.73436$, $w_{27} = 0.79179$,
 $w_{28} = 0.84397$.

Figura 7.11: Experimentos usando a base de dados BDSHAPES.

Tabela 7.4: Mean Average Precision dos experimentos: (1) representação normalizada. (2) representação das características utilizando conjuntos difusos sem ponderação no cálculo da similaridade. (3) representação das características utilizando conjuntos difusos com ponderação no cálculo da similaridade. Também é mostrado o ganho relativo da utilização da representação proposta em relação a representação normalizada.

	experimento 1	experimento 2	experimento 3	ganho relativo
<i>IRIS</i>				
Distância Euclideana	87.1013	87.6764	88.7739	1.92 %
Distância City block	87.2963	87.5797	88.7467	1.66 %
Similaridade Cosseno	71.9565	88.121	89.5135	24.39 %
Equality Index	84.8114	86.0313	87.3069	2.94 %
Distância baseada em norma	79.2103	86.7305	87.9636	11.05 %
Similaridade baseado em lógica	87.2977	86.0309	87.3069	0.01 %
Similaridade proposta por Wang	87.2977	86.0309	87.3069	0.01 %
<i>WDBC</i>				
Distância Euclideana	82.5198	85.1802	85.5343	3.65 %
Distância City block	83.1287	85.1845	85.5387	2.89 %
Similaridade Cosseno	72.8988	84.0827	84.4203	15.80 %
Equality Index	84.6979	85.1735	85.5279	0.98 %
Distância baseada em norma	68.1639	84.7797	85.1133	24.86 %
Similaridade baseado em lógica	83.1287	85.1845	85.5387	2.89 %
Similaridade proposta por Wang	83.1287	85.1845	85.5387	2.89 %
<i>BD111</i>				
Distância Euclideana	79.3428	80.6037	80.5373	1.58 %
Distância City block	79.6363	80.7512	80.7375	1.40 %
Similaridade Cosseno	67.1947	78.3262	78.3563	16.61 %
Equality Index	80.3026	80.4433	80.4292	0.17 %
Distância baseada em norma	69.873	80.552	80.5317	15.28 %
Similaridade baseado em lógica	79.6363	80.5949	80.5676	1.20 %
Similaridade proposta por Wang	79.6363	80.5949	80.5676	1.20 %
<i>BDFACES</i>				
Distância Euclideana	57.2131	50.4044	50.2520	-
Distância City block	56.8839	51.0889	51.0056	-
Similaridade Cosseno	56.831	44.9963	45.0376	-
Equality Index	57.4984	49.8169	49.6582	-
Distância baseada em norma	55.8125	49.0868	48.9427	-
Similaridade baseado em lógica	56.8839	50.3847	50.2200	-
Similaridade proposta por Wang	56.8839	50.3847	50.2200	-
<i>BDSHAPES</i>				
Distância Euclideana	49.6047	44.5502	44.5462	-
Distância City block	47.5368	44.2508	44.2247	-
Similaridade Cosseno	50.3567	43.7284	43.7557	-
Equality Index	44.7483	44.1501	44.033	-
Distância baseada em norma	49.8873	44.0359	43.9912	-
Similaridade baseado em lógica	47.5368	44.53	44.4293	-
Similaridade proposta por Wang	47.5368	44.53	44.4293	-

Tabela 7.6: Mean Average Precision utilizando diferentes partições difusas.

base de dados	MANGPD	PDU	PDK	HFP
IRIS	88.7739	88.1	88.5659	88.2825
WDBC	85.5343	83.0468	83.9013	78.1065
BD111	80.6037	78.4788	81.3481	78.2161
BDFACES	50.2520	55.2884	54.7695	55.6011
BDSHAPES	44.5462	45.0924	46.2355	45.3371

7.5 Considerações Finais do Capítulo

Neste capítulo foi apresentada a avaliação dos resultados obtidos nos experimentos.

Basicamente, foram feitas comparações entre a representação normalizada e a representação proposta. Utilizou-se bases de dados sintéticas e bases de dados reais, sendo que a representação proposta apresentou os melhores resultados em todas as bases de dados sintéticas e também nas bases de dados reais IRIS, WDBC e BD111, independentemente do operador de similaridade.

Na comparação entre os métodos para geração de partição difusa, o método proposto no trabalho obteve os melhores resultados em todas as bases de dados sintéticas e o melhor resultado em duas das bases de dados reais. Notou-se que o método proposto não obtém bons resultados em bases de dados em que o resultado do agrupamento possui valores altos de entropia.

Capítulo 8

Conclusão

Várias medidas de similaridade têm sido descritas na literatura. A qualidade dos resultados em sistemas CBIR depende da escolha da métrica a ser utilizada na busca por similaridade. Neste trabalho é proposta uma nova representação para os vetores de característica que torna a recuperação dos dados menos sensível à métrica escolhida. A representação proposta é baseada em conjuntos difusos obtidos a partir das partições difusas. A análise dos experimentos computacionais mostraram a eficácia da nossa proposta.

Também foi proposta neste trabalho uma nova estratégia para a obtenção de partição difusa utilizando o Fuzzy C-Means. A partição difusa obtida pode empregar funções de pertinência com diferentes formas, dependendo do agrupamento resultante. O ponto de interseção entre duas funções de pertinência adjacentes não é previamente fixado como em outros trabalhos relacionados na literatura. Os experimentos realizados até o momento indicam que ambos os métodos propostos neste trabalho apresentaram bons resultados quando comparados a outras abordagens similares.

As principais contribuições deste trabalho são:

- desenvolvimento de uma nova maneira de representar as características e consequentemente como calcular a similaridade utilizando esta nova representação;
- um novo método automático e não supervisionado para a geração da partição difusa a partir da base de dados.

Há algumas propostas de trabalhos futuros. Como o método proposto não obteve bons resultados nas bases de dados em que o resultado do agrupamento apresentou um valor de entropia alto, fica em aberto para trabalhos futuros a utilização de outras técnicas de agrupamento que consigam minimizar o valor da entropia, possibilitando assim alcançar melhores resultados nestas bases de dados.

Outra proposta para trabalhos futuros é utilizar a partição difusa gerada pelo método proposto, visando a geração de regras difusas em sistemas de classificação.

Referências Bibliográficas

- [Balopoulos et al. 2007] Balopoulos, V., Hatzimichailidis, A. G., e Papadopoulos, B. K. (2007). Distance and similarity measures for fuzzy operators. *Information Sciences*, 177(11):2336 – 2348.
- [Bezdek 1981] Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA.
- [Bustince et al. 2007] Bustince, H., Pagola, M., e Barrenechea, E. (2007). Construction of fuzzy indices from fuzzy DI-subsethood measures: Application to the global comparison of images. *Information Sciences*, 177(3):906 – 929.
- [Campello e Hruschka 2006] Campello, R. e Hruschka, E. (2006). A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems*, 177:2858 – 2875.
- [Candan et al. 2000] Candan, K. S., Li, W.-S., e Priya, M. L. (2000). Similarity-based ranking and query processing in multimedia databases. *Data Knowl. Eng.*, 35(3):259–298.
- [Chaudhuri e Rosenfeld 1999] Chaudhuri, B. B. e Rosenfeld, A. (1999). A modified Hausdorff distance between fuzzy sets. *Inf. Sci.*, 118(1-4):159–171.
- [Chiu 1994] Chiu, S. L. (1994). Fuzzy model identification based on cluster estimation. *Intelligent and Fuzzy Systems*, 2:267 – 278.
- [Choras et al. 2007] Choras, R., Andrysiak, T., e Choras, M. (2007). Integrated color, texture and shape information for content-based image retrieval. *Pattern Analysis Applications*, 10:333–343.
- [Datta et al. 2008] Datta, R., Joshi, D., Li, J., e Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60.
- [DBGen] DBGen. Database Generator - DBGen. Grupo de Bases de Dados e de Imagens (GBDI) do ICMC/USP. URL=<http://www.gbdi.icmc.usp.br/download>.
- [Dunn 1973] Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, 3:32 – 57.
- [Faria et al. 2005] Faria, E., Guliato, D., e de Sousa Santos, J. C. (2005). Segmentation and centromere locating methods applied to fish chromosomes images. *Brazilian Symposium on Bioinformatics Proceedings*, pp. 181–189.

- [Flickner et al. 1995] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., e Yanker, P. (1995). Query by Image and Video Content: The QBIC System. *Computer*, 28:23–32.
- [Guillaume e Charnomordic 2004] Guillaume, S. e Charnomordic, B. (2004). Generating an interpretable family of fuzzy partitions. *IEEE Transactions on Fuzzy Systems*, 12(3):324 – 335.
- [Guillaume et al. 2001] Guillaume, S., Charnomordic, B., e Titli, A. (2001). A distance metric suitable for fuzzy partitioning. *The 10th IEEE International Conference on Fuzzy Systems*, 1:264 – 267.
- [Guliatto et al. 2008] Guliatto, D., de Carvalho, J., Rangayyan, R., e Santiago, S. (2008). Feature Extraction from a Signature Based on the Turning Angle Function for the Classification of Breast Tumors. *Journal of Digital Imaging*, 21:129–144. 10.1007/s10278-007-9069-9.
- [Guliatto et al. 2009] Guliatto, D., de Melo, E. V., Rangayyan, R. M., e Soares, R. C. (2009). POSTGRESQL-IE: An Image-handling Extension for PostgreSQL. *Journal of Digital Imaging*, 22(2):149–165.
- [Guliatto et al. 1998] Guliatto, D., RANGAYYAN, R. M., CARNIELLI, W., ZUFFO, J., e DESAUTELS, J. (1998). Segmentation of breast tumors in mammograms by fuzzy region growing. *Proc. 20th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society, Hong Kong*, pp. 1002–1004.
- [Guliatto e Santos 2009] Guliatto, D. e Santos, J. C. S. (2009). Granular Computing and Rough Sets to Generate Fuzzy Rules. *Proceedings of the 6th International Conference on Image Analysis and Recognition*, pp. 317–326.
- [Hathaway e Bezdek 2001] Hathaway, R. J. e Bezdek, J. C. (2001). Fuzzy c-means clustering of incomplete data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 31(5):735 – 744.
- [Hjaltason e Samet 2003] Hjaltason, G. R. e Samet, H. (2003). Index-driven similarity search in metric spaces. *ACM Transactions on Database Systems*, 28:2003.
- [Hong e Chen 1999] Hong, T.-P. e Chen, J.-B. (1999). Finding relevant attributes and membership functions. *Fuzzy Sets and Systems*, 103(3):389 – 404.
- [Hong e Lee 1996] Hong, T.-P. e Lee, C.-Y. (1996). Induction of fuzzy rules and membership functions from training examples. *Fuzzy Sets and Systems*, 84(1):33 – 47.
- [Ishibuchi e Nakashima 2000] Ishibuchi, H. e Nakashima, T. (2000). Linguistic rule extraction by genetics-based machine learning. *Proc. of Genetic and Evolutionary Computation Conference*, pp. 195 – 202.
- [Ishibuchi et al. 1992] Ishibuchi, H., Nozaki, K., e Tanaka, H. (1992). Distributed representation of fuzzy rules and its application to pattern classification. *Fuzzy Sets Syst.*, 52(1):21–32.

- [Kbir et al. 2000] Kbir, M. A., Benkirane, H., Maalmi, K., e Benslimane, R. (2000). Hierarchical fuzzy partition for pattern classification with fuzzy if-then rules. *Pattern Recognition Letters*, 21(6-7):503 – 509.
- [Klir e Yuan 1995] Klir, G. J. e Yuan, B. (1995). *Fuzzy Sets and Fuzzy Logic, Theory and Applications*. Prentice Hall, New Jersey.
- [Kuncheva 2000] Kuncheva, L. I. (2000). Fuzzy Classifier Design. *Physica-Verlag, Heidelberg*.
- [Luukka 2007] Luukka, P. (2007). Similarity classifier using similarity measure derived from Yu's norms in classification of medical data sets. *Computers in Biology and Medicine* 37, pp. 1133–1140.
- [MacQueen 1967] MacQueen, J. B. (1967). Some Methods for Classification and Analysis of MultiVariate Observations. In Cam, L. M. L. e Neyman, J. (editores), *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 281–297. University of California Press.
- [Manning et al. 2008] Manning, C. D., Raghavan, P., e Schütze, H. (2008). *An Introduction to Information Retrieval*. Cambridge University Press.
- [Medasani et al. 1998] Medasani, S., Kim, J., e Krishnapuram, R. (1998). An overview of membership function generation techniques for pattern recognition. *International Journal of Approximate Reasoning*, 19(3-4):391 – 417.
- [Mendel 1995] Mendel, J. (1995). Fuzzy logic systems for engineering: a tutorial. *Proceedings of the IEEE*, 83(3):345 –377.
- [MIAS] MIAS. The Mammographic Image Analysis Society digital mammogram database. URL=<http://http://peipa.essex.ac.uk/info/mias.html>.
- [Murata et al. 1998] Murata, T., Ishibuchi, H., e Gen, M. (1998). Adjusting Fuzzy Partitions by Genetic Algorithms and Histograms for Pattern. *IEEE International Conference on Evolutionary Computation*, pp. 9 – 14.
- [Nordstrøm et al. 2004] Nordstrøm, M. M., Larsen, M., Sierakowski, J., e Stegmann, M. B. (2004). The IMM Face Database - An Annotated Dataset of 240 Face Images.
- [Ovchinnikov 1991] Ovchinnikov, S. (1991). Similarity relations, fuzzy partitions, and fuzzy orderings. *Fuzzy Sets Syst.*, 40(1):107–126.
- [Ovchinnikov 1981] Ovchinnikov, S. V. (1981). Structure of fuzzy binary relations. *Fuzzy Sets and Systems*, 6(2):169 – 195.
- [Pal e Pal 1993] Pal, N. R. e Pal, S. K. (1993). A review on image segmentation techniques. *Pattern Recognition* 26, pp. 1277–1294.
- [ScreenTest] ScreenTest. Alberta Program for the Early Detection of Breast Cancer 2001/03. URL=<http://www.cancerboard.ab.ca/screentest>.
- [Shapes] Shapes. ImageProcessingPlace.com - MPEG7 CE Shape-1 Part B. URL=http://www.imageprocessingplace.com/downloads_V3/root_downloads/image_databases/.

- [Smith e Chang 1997] Smith, J. R. e Chang, S.-F. (1997). Visualseek: a fully automated content-based image query system. In *Proceedings of the fourth ACM international conference on multimedia*, pp. 87–98, Boston, Massachusetts, United States.
- [Steinbach et al. 2006] Steinbach, M., Tan, P.-N., e Kumar, V. (2006). *Introduction to Data Mining*. Pearson Addison Wesley.
- [Tobias e Seara 2002] Tobias, O. e Seara, R. (2002). Image segmentation by histogram thresholding using fuzzy sets.
- [UCI | UCI. UC Irvine Machine Learning Repository. URL=<http://archive.ics.uci.edu/ml/>.
- [Veltkamp e Tanase 2002] Veltkamp, R. C. e Tanase, M. (2002). Content-based Image and Video Retrieval. In *Proceedings of the Fifth IASTED International Conference on Biomedical Engineering, Innsbruck, Austria, Austria*.
- [Wang 1997] Wang, W.-J. (1997). New similarity measures on fuzzy sets and on elements. *Fuzzy Sets and Systems*, 85(3):305 – 309.
- [Wu e Mendel 2008] Wu, D. e Mendel, J. M. (2008). A vector similarity measure for linguistic approximation: Interval type-2 and type-1 fuzzy sets. *Information Sciences*, 178(2):381 – 402.
- [Xuecheng 1992] Xuecheng, L. (1992). Entropy, distance measure and similarity measure of fuzzy sets and their relations. *Fuzzy Sets and Systems*, 52(3):305 – 318.
- [Zadeh 1965] Zadeh, L. (1965). Fuzzy sets. *Information and Control*, vol 8, Issue 3, pp. 338–353.
- [Zadeh 1971] Zadeh, L. A. (1971). Similarity relations and fuzzy orderings. *Inf. Sci.*, 3(2):177–200.
- [Zhang e Fu 2006] Zhang, C. e Fu, H. (2006). Similarity measures on three kinds of fuzzy sets. *Pattern Recognition Letters*, 27(12):1307 – 1317.
- [Zwick et al. 1987] Zwick, R., Carlstein, E., e Budescu, D. V. (1987). Measures of similarity among fuzzy concepts: a comparative analysis. *International Journal of Approximate Reasoning* 1, pp. 221 – 242.