Novos Métodos de Classificação Nebulosa e de Validação de Categorias e suas Aplicações a Problemas de Reconhecimento de Padrões

Cláudia Rita de Franco

Universidade Federal do Rio de Janeiro
Curso de Mestrado

Adriano Joaquim de Oliveira Cruz
Ph.D.

Rio de Janeiro
2002

Novos Métodos de Classificação Nebulosa e de Validação de Categorias e suas Aplicações a Problemas de Reconhecimento de Padrões

Cláudia Rita de Franco

Dissertação submetida ao corpo docente do DCC/IM e NCE/UFRJ, Universidade Federal do Rio de Janeiro – UFRJ, como parte dos requisitos necessários à obtenção do grau de Mestre.

Aprovada por:

Prof. _____– Orientador
Adriano Joaquim de Oliveira Cruz – Ph.D.

Prof. _____
Marco Aurélio Cavalcanti Pacheco – Ph.D.

Prof. _____
Antonio Carlos Gay Thomé – Ph.D.

Rio de Janeiro
2002

Franco, Cláudia Rita de

Novos Métodos de Classificação Nebulosa e Validação de Categorias e suas Aplicações a Problemas de Reconhecimento de Padrões / Cláudia Rita de Franco. Rio de Janeiro: UFRJ/IM-NCE, 2002.

XXI, n p.133; il.

Dissertação (Mestrado) – Universidade Federal do Rio de Janeiro, IM/NCE, 2002.

1. Reconhecimento Estatístico de Padrões – Tese. 2. Categorização Nebulosa – Tese. 3. Validação de Categorias – Tese. I. Título. II. Tese (Mestr. – UFRJ/IM-NCE). III. Autor.

Ao Leonardo, que mesmo navegando no mesmo mar revolto de incertezas sempre me apoiou incondicionalmente.

Ao meu pai, que partiu no inicio desta minha caminhada...

À minha mãe, por todos os momentos de preocupação ao ver o cansaço no meu rosto.

AGRADECIMENTOS

Ao Prof. Adriano, que me orienta desde a graduação, por confiar nas minhas decisões e por não duvidar da minha capacidade, o que fiz por vezes. Finalmente, obrigado por ter sido além de professor, um grande amigo.

Ao NCE e à FAPERJ, pelo apoio financeiro dedicado à pesquisa realizada neste trabalho.

Ao Chakhra e ao Qui-Gon, por me mostrarem a alegria da vida toda vez que eu olhava os seus rostinhos.

À Norma e ao Leo, meus segundos pais, que sempre estiveram presentes, me ajudando e apoiando, mesmo quando o caminho parecia tão escuro.

À minha mãe, que sempre esteve ao meu lado, me ensinando a pensar positivamente em todas as situações da vida.

Ao meu pai, que mesmo não estando presente fisicamente, sempre confiou e me apoiou nas minhas escolhas e decisões. Pai, você ficará no meu coração para sempre...

Ao Leonardo, por me lembrar que existem momentos importantes na vida que devem ser vividos intensamente e por seus “empurrões” nos momentos de desânimo.

RESUMO

FRANCO, Cláudia Rita de. **Categorização Nebulosa e Validação de Categorias Aplicadas a Problemas de Reconhecimento de Padrões**. Orientador: Adriano Joaquim de Oliveira Cruz. Rio de Janeiro: UFRJ/DCC-NCE, 2002. Dissertação (Mestrado em Ciência da Computação).

A Categorização tem um importante papel em muitas áreas de pesquisa, especialmente nas que envolvem problemas de Reconhecimento de Padrões. Geralmente, nos problemas do mundo real, o número de classes é desconhecido, sendo necessário ter critérios que identifiquem a melhor escolha de categorias. Nesta dissertação é proposta uma extensão para o Discriminante Linear de Fisher, o EFLD, que pode ser aplicado a partições rígidas e nebulosas, e um método de validação rápido e eficiente que mede a compacidade e a separação das partições produzidas por qualquer método de categorização nebuloso ou rígido. As simulações realizadas indicam que ela é uma medida eficiente e rápida mesmo quando a sobreposição das categorias é alta.

Baseado nas vantagens de dois métodos de categorização muito conhecidos e na medida de validação proposta, um Sistema Estatístico Não-Paramétrico de Reconhecimento de Padrões é proposto nesta tese. Este algoritmo é avaliado através do problema de Reconhecimento de Dígitos Cursivos.

ABSTRACT

FRANCO, Cláudia Rita de. **Categorização Nebulosa e Validação de Categorias Aplicadas a Problemas de Reconhecimento de Padrões**. Orientador: Adriano Joaquim de Oliveira Cruz. Rio de Janeiro: UFRJ/DCC-NCE, 2002. Dissertação (Mestrado em Ciência da Computação).

Cluster analysis has an important role in many research areas, especially those involving problems of pattern recognition. Generally, in real world problems, the number of classes is unknown in advance, being necessary to have criterions to identify the best choice of clusters. In this dissertation is proposed an extension to Fisher Linear Discriminant, the EFLD, that can be applied to fuzzy and crisp partitions and a new fast and efficient validity method that measures the compactness and separation of partitions produced by any fuzzy or crisp clustering method. The simulations performed indicate that it's an efficient and fast measure even when the overlapping between clusters is high.

Based on the advantages of two very known fuzzy clustering methods and in the proposed validity measure, a Non-Parametric Statistical Pattern Recognition System is proposed in this thesis. This algorithm is evaluated to the problem of Cursive Digits Recognition.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

O ser humano é munido da capacidade de reconhecer e classificar padrões, o que o leva a ter uma percepção única do mundo. É através da observação dos acontecimentos que é cunhada a relação entre o que é experimentado e o que é compreendido.

Quanto mais o ser humano experimenta, mais ele é capaz de compreender e classificar o que reconheceu em estruturas de acordo com a similaridade de suas características e padrões.

Como ciência, o *Reconhecimento de Padrões* vem sendo estudado desde o início de 1950, dado o avanço da tecnologia que levou ao advento dos computadores. Ele pode ser definido como "o processo de identificar estruturas em dados através de comparações com estruturas conhecidas". (ROSS, 1997, p. 411)

O *Reconhecimento de Padrões* também pode ser definido como uma área de pesquisa e desenvolvimento de sistemas que objetivam reconhecer padrões em dados. Esta é uma área grande que engloba várias subáreas como análise de características (feature analysis), estimação de erro, análise de discriminante, categorização (cluster analysis), dentre outras.

Esta técnica vem sendo aplicada em diversas áreas como:

- Reconhecimento de caracteres manuscritos e de palavras;
- Diagnósticos médicos, triagem automatizada, classificação de imagens de raios-X, eletrocardiogramas, eletroencefalogramas;
- Diagnóstico de máquinas e inspeção industrial;
- Identificação pessoal, como reconhecimento de fala e identificação de voz, reconhecimento de rostos humanos e reconhecimento de impressão digital;
- Classificação de ondas sísmicas, geoprocessamento;
- Identificação de alvos, dentre outras.

1.1 ANÁLISE DE CARACTERÍSTICAS

A *Análise de Características* consiste em um conjunto de métodos que transformam dados brutos em informações que podem ser analisadas e comparadas.

Cada dado ou amostra é então representado por um vetor multidimensional de dados, onde cada elemento é chamado de *característica* (feature). Este vetor é denominado *vetor de características*.

A *Análise de Características* é constituída de três etapas. A primeira delas, a *Identificação de Características* (Feature Nomination), é o processo de gerar as características das amostras. As características são geradas por processos físicos, como medidas extraídas de sensores, ou por processos matemáticos.

A segunda etapa é a *Seleção de Características* (Feature Selection), onde é escolhido o subconjunto *s*, com $s < p$, do conjunto original com *p* características que melhor representa as amostras.

É na etapa de *Extração de Características* (Feature Extraction) que é realizado o processo de reduzir a dimensionalidade dos vetores de características, preservando ou aprimorando as informações contidas nestes.

O objetivo desta etapa é eliminar as características inúteis ou redundantes, facilitando a análise do problema sem comprometê-la. O método mais utilizado é a combinação linear das medidas iniciais, que pode ser obtida através de processos como a *Análise dos Componentes Principais* e o *Discriminante Linear de Fisher*. (KUPAC, 2000) (BISHOP, 1995; DUDA, 1973)

1.2 CATEGORIZAÇÃO

Categorização é o processo de particionar um conjunto de amostras em *subconjuntos* ou *categorias* cujos dados são similares entre si através de suas características.

Categorias são conjuntos cujas entidades apresentam semelhanças entre as características que são utilizadas matematicamente para representá-las enquanto que *classes* são conjuntos de entidades que compartilham características semelhantes segundo um critério do mundo real.

Os *métodos de categorização* apresentados nesta dissertação se baseiam em métricas de distância que associam os dados através de sua disposição espacial.

Estes métodos partem da premissa de que dados que estão próximos pertencem a uma mesma categoria, enquanto que dados distantes pertencem a categorias diferentes. Neste caso, os dados são projetados no espaço $\Re^p$ a partir de suas *p* características, que correspondem às suas coordenadas neste espaço.

Os *métodos de categorização* podem ser *rígidos* (hard ou crisp) ou *nebulosos* (fuzzy ou soft). Nos *métodos rígidos*, cada amostra é associada a uma única categoria gerada. Nos *métodos nebulosos*, cada amostra é associada a todas as categorias geradas conforme sua afinidade com cada categoria.

Os métodos de *categorização* podem ser empregados em reconhecimento de padrões, construção de taxonomias em biologia e em outras áreas, classificação de documentos por tipo de informações, análise de imagens, dentre outros.

1.2.1 VALIDAÇÃO DE CATEGORIAS

Uma dificuldade enfrentada na *categorização* é a avaliação do número de categorias em que se deve particionar um conjunto de amostras.

Quando as classes de um problema são conhecidas, considera-se que a melhor partição do espaço amostral também é conhecida, sendo o número de categorias ideal igual ao número de classes do problema.

Porém existem muitos casos em que o número de classes de um problema é desconhecido. Isto ocorre quando os dados são *não rotulados*, havendo a necessidade de identificar o número de categorias que expresse a melhor estruturação dos dados.

O estudo de métricas que identificam o número ótimo de partições de um conjunto amostral é chamado de *Validação de Categorias* (Cluster Validity).

1.3 RECONHECIMENTO ESTATÍSTICO DE PADRÕES

O *Reconhecimento de Padrões* associado à *Categorização* é chamado de *Sistema Estatístico Não-Paramétrico de Reconhecimento de Padrões* (Non-Parametric Statistical Pattern Recognition). (HORIKAWA, 1997)

Os *Sistemas Estatísticos de Reconhecimento de Padrões* se baseiam em modelos matemáticos. As medidas matemáticas utilizadas para avaliar a similaridade dos dados com as classes do problema avaliam as propriedades geradoras das características dos dados que os tornam fisicamente similares, como por exemplo a distância entre as características das amostras.

Estes sistemas podem ser *paramétricos* ou *não-paramétricos*. Nos sistemas *paramétricos*, o tipo de distribuição probabilística das amostras é previamente definido por um modelo estatístico. Nos sistemas *não-paramétricos*, o reconhecimento de padrões é realizado a partir das características das amostras, sem que haja um conhecimento prévio da distribuição destas.

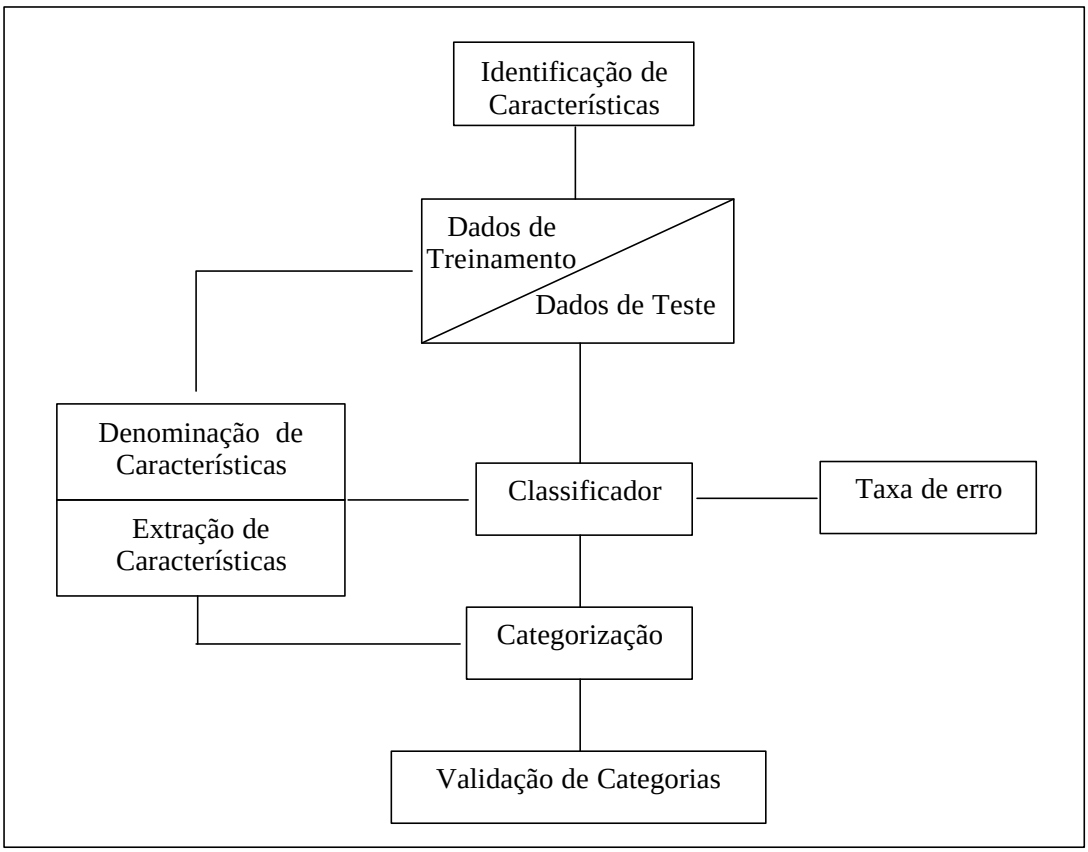

Figura 1.1 – Representação do Sistema Estatístico Não-Paramétrico de Reconhecimento de Padrões (BEZDEK, 1996)

A *categorização* tem um papel crucial nestes sistemas, pois ela é responsável por encontrar as estruturas que formam os dados e agrupá-los em categorias. Em um sistema

não supervisionado, ou seja, em problemas cujas classes não são conhecidas, as categorias geradas corresponderão às classes do problema.

O *Reconhecimento de Padrões* é responsável por associar dados novos às classes geradas pelos métodos de categorização. Os dados utilizados pelo sistema são divididos em dados de treinamento e teste. Os componentes de um sistema de *Sistema Estatístico Não-Paramétrico de Reconhecimento de Padrões* segundo Bezdek são ilustrados na Figura 1.1.

O módulo *classificador* classifica os dados de testes e os dados novos utilizando as categorias geradas no módulo de *categorização*. As categorias são geradas pela categorização dos dados de treinamento e são avaliadas pelo modulo de *validação de categorias*.

A associação entre a *Categorização Nebulosa* e o *Reconhecimento de Padrões* foi consolidada durante a década de 1960 e é muito vasta na literatura atual, o que é natural dado que, no mundo real, as classes de um problema geralmente não têm fronteiras bem definidas.

1.4 OBJETIVOS DO TRABALHO DE PESQUISA

Este trabalho visou o estudo dos *métodos de categorização nebulosos* e de suas aplicações a problemas de *Reconhecimento de Padrões*.

Um dos aspectos mais importantes da *categorização* é o estudo do número ideal de categorias em que deve ser particionado um espaço amostral.

Este estudo culminou na extensão de uma medida de validação existente, capacitando-a a avaliar espaços nebulosos, além dos rígidos, e na proposta de uma nova *medida de validação* rápida e eficiente capaz de analisar partições nebulosas e rígidas.

As duas medidas foram submetidas a testes para avaliar seu desempenho.

A nova medida de validação mostrou-se capaz de analisar de forma eficiente o espaço amostral, avaliando como ideal o número de categorias esperado, mesmo nos casos em que a sobreposição das amostras é alta, que é a grande responsável pela falha das outras medidas de validação.

O levantamento das características dos métodos de categorização levou à proposta de um *Sistema de Reconhecimento Estatístico de Padrões*, que combina as funcionalidades de dois métodos de categorização nebulosos vastamente conhecidos e da medida de validação proposta.

Finalmente, o sistema proposto foi aplicado ao problema de *Reconhecimento de Dígitos Manuscritos.*

1.5 PRINCIPAIS MOTIVAÇÕES

A área de Inteligência Computacional é relativamente nova, oferecendo grandes desafios, o que a torna uma área de pesquisa fascinante para um trabalho de mestrado. Porém é o seu aspecto interdisciplinar que mais atrai os pesquisadores, sendo possível aplicar suas técnicas a diversas áreas, como medicina, geoprocessamento, processos industriais, biologia, etc.

O processo de Reconhecimento de Padrões faz parte do estudo da percepção. É a partir da percepção que o ser humano estabelece seu relacionamento com o mundo e desenvolve a sua inteligência. Fazendo parte deste processo, encontra-se a *categorização*, que é uma técnica interessante, pois tem um papel muito importante nos problemas de reconhecimento de padrões.

É ela quem descobre as estruturas dos elementos. E a maior motivação é que tudo que nos rodeia tem uma estrutura, algumas ainda desconhecidas para nós. Criar sistemas automatizados capazes de reconhecer e classificar estas estruturas é uma forma de automatizar a percepção, levando-nos a entender fenômenos desconhecidos.

Finalmente, foi a possibilidade de aprender e combinar as funcionalidades das técnicas de categorização para criar um sistema de reconhecimento de padrões o maior motivo de motivação para o desenvolvimento do trabalho mostrado nesta dissertação.

1.6 ORGANIZAÇÃO DA DISSERTAÇÃO

No capítulo 2, são apresentados diversos métodos de categorização rígidos e nebulosos, que são amplamente utilizados na literatura, e seus algoritmos, com uma

visão geral de suas aplicações. Também são apresentadas as medidas de validações pesquisadas neste trabalho e o Discriminante Linear de Fisher.

O capítulo 3 apresenta a proposta do Discriminante Linear de Fisher Estendido, que se torna qualificado para analisar partições nebulosas, além das rígidas. Este capítulo também apresenta a proposta de uma nova medida de validação, adequada para avaliar tanto partições rígidas como nebulosas.

Neste capítulo também são mostrados experimentos onde é avaliado o desempenho destas duas medidas.

O capítulo 4 apresenta a descrição detalhada e a avaliação do Sistema Estatístico Não-Paramétrico de Reconhecimento de Padrões proposto nesta dissertação.

No capítulo 5, o sistema proposto é aplicado ao problema de Reconhecimento de Dígitos Manuscritos e tem seus resultados comparados com outros métodos de categorização nebulosos.

O capítulo 6 apresenta as conclusões e as dificuldades encontradas na elaboração dos estudos. Também são apresentadas algumas sugestões para trabalhos futuros.

2 CATEGORIZAÇÃO E CLASSIFICAÇÃO

A *Categorização* (Cluster Analysis) é uma técnica utilizada para descobrir estruturas em conjuntos de dados a fim de dividi-los em subconjuntos denominados *categorias* (clusters) cujos elementos tem aspectos similares entre si (Figura 2.1).

As categorias, também chamadas de *agrupamentos* ou *partições*, agrupam dados que são similares entre si através de seus *atributos*, *padrões* ou *características*. A similaridade entre os pares de pontos da amostra é estimada a partir de uma medida estipulada.

Como a representação matemática dos dados geralmente engloba apenas um subconjunto das características das entidades do mundo real, a categorização só poderá buscar semelhanças neste subconjunto de características.

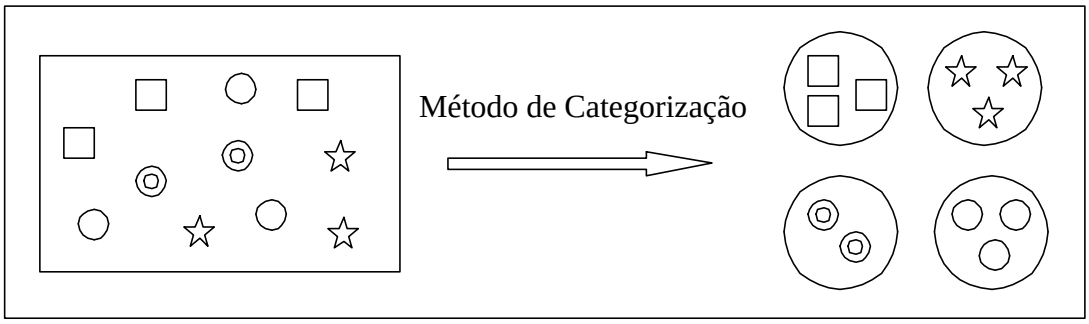

Figura 2.1 – Categorização

Conseqüentemente, as semelhanças encontradas podem não corresponder às principais semelhanças existentes no mundo real, produzindo categorias que não representam as classes do problema.

A métrica mais simples e mais utilizada nos métodos de categorização é o cálculo da *distância Euclidiana*, (eq. 2-4), entre as *características* (features) ou *coordenadas* dos pontos da amostra.

Ao utilizar esta medida, presume-se que se dois pontos pertencem à mesma categoria, a distância entre eles deve ser menor que a distância entre dois pontos que pertençam a categorias diferentes.

Os métodos de categorização associam as amostras de um conjunto às categorias geradas e não às classes originais do problema, sendo necessário criar um método para associar as amostras às classes do problema.

Os métodos de categorização são classificados em dois tipos: *hierárquicos* e *não-hierárquicos*.

Nos métodos *não-hierárquicos*, os dados são distribuídos pelo número de categorias desejadas e um critério é otimizado. Um dos critérios empregados é a minimização da variação interna das categorias obedecendo a um número escolhido de partições.

Nos métodos *hierárquicos* são utilizados dois tipos de abordagens. Na primeira abordagem, cada ponto do espaço amostral é um centro de categoria. Nos próximos passos, cada dois pontos com distância mínima entre si são sucessivamente fundidos em uma única categoria até que o número desejado de categorias seja atingido.

Na segunda abordagem, todos as amostras são alocadas em uma única categoria e um critério é usado para dividi-la no número de categorias esperado.

Os métodos de categorização são *não-supervisionados*, pois ao categorizar um espaço amostral, mesmo para dados rotulados, os métodos não consideram em seus algoritmos a classificação esperada de cada amostra. Estes métodos só consideram as características físicas utilizadas na representação das amostras.

Os métodos de categorização também podem ser *rígidos* (crisp) ou *nebulosos* (fuzzy). Nos métodos rígidos, cada ponto da amostra pertence a uma e somente uma categoria.

Nos métodos nebulosos, cada amostra pertence a todos as categorias com diferentes graus de afinidade. Estes graus são denominados *graus de inclusão* (membership).

Outro aspecto da categorização é o número ideal de categorias em que um espaço amostral deve ser particionado. Para dados *rotulados* o valor ideal do número de categorias é conhecido. Mas quando os dados são *não rotulados*, é necessário definir uma métrica que identifique o valor ótimo. Este problema é conhecido como *Validação de Categorias* (Cluster Validity).

A *Classificação* é uma técnica que associa amostras a classes previamente conhecidas. Os métodos de *Categorização* podem ser associados à *Classificação* com a finalidade de encontrar padrões nos dados que gerem as categorias que funcionarão como classes na técnica de classificação.

Muitos autores como Timothy J. Ross misturam os conceitos de categorização e de classificação, chegando a fundir os dois. Ao explicar os métodos de categorização nebulosos, o autor se refere a estes como métodos de classificação nebulosos. (ROSS, 1997)

Assim como os métodos de categorização, existem métodos de classificação *nebulosos* e *rígidos*. Além disto, os métodos de classificação podem ser considerados *supervisionados* ou *não supervisionados*.

A rede neural *MLP* é um exemplo de *método de classificação supervisionado rígido*, pois dado a entrada e a saída desejadas, ela se configura internamente através do seu treinamento para gerar a saída esperada associando assim a amostra fornecida a uma classe.

O *K-NN* (K-Nearest Neighbour Algorithm) assim como sua versão nebulosa, o *K-NN nebuloso* (Fuzzy K-NN Classifier), são exemplos de *métodos de classificação rígido e nebuloso*, respectivamente.

Segundo Pavlidis (2001, p.7), o método "*K-NN* é o único entre os métodos supervisionados de aprendizado no qual não há o passo de treinamento".

Porém, segundo Braga, o conceito de *aprendizado supervisionado* envolve um conjunto de treinamento que possua pares de entrada e saída desejados, o que não acontece no método *K-NN*, dado que não existe uma fase de treinamento no algoritmo e que só são considerados os dados de entrada ao classificar amostras, visto que o método não tem conhecimento dos dados de saída esperados. (BRAGA, 1998)

Segundo este conceito, os métodos *K-NN rígido* e *K-NN nebuloso* são *métodos de classificação não supervisionados*, o que será considerado neste trabalho.

Na seção 2.1 deste capítulo será apresentado o método de categorização rígido não-hierárquico *k-Means* e o método de classificação rígido *K-NN*.

Na seção 2.2 será apresentada a versão nebulosa do método *k-Means*, o *Fuzzy c-Means (FCM)* e os métodos *Gustafson-Kessel (GK)* e *Gath-Geva (GG)*, derivados do *FCM*. Também será apresentada a *rede neural FKCN*, inspirada no método *FCM* e na rede Kohonen (BRAGA, 1998), e a versão nebulosa do método de classificação rígido *K-NN*, o *K-NN nebuloso*.

Na seção 2.3 serão apresentadas as *medidas de validação* de partições nebulosas mais utilizadas na literatura e finalmente na seção 2.3.10 será apresentado o *Discriminante Linear de Fisher*, que também pode ser empregado como um validador de categorias rígidas.

2.1 MÉTODOS RÍGIDOS DE CATEGORIZAÇÃO E CLASSIFICAÇÃO

2.1.1 MÉTODO K-MEANS

O *k-Means* é um método de categorização *não-hierárquico rígido*, criado para agrupar dados não rotulados em *categorias*. É um dos métodos rígidos mais utilizados por não impor restrições ao conjunto de amostras, podendo ser aplicado a qualquer quantidade de dados. (ROSS, 1997)

Para um conjunto $X = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ de n amostras *não-rotuladas* do espaço $\Re^p$, onde as classes do problema não são conhecidas, o método *k-Means* associa cada amostra a uma única *categoria* ou *partição*.

As categorias geradas assumem a forma de *hiperesferas* do $\Re^p$ de mesmo tamanho, que são caracterizadas pelos seus centros.

Para o conjunto $A = \{a_1, \ldots, a_c\}$ de c partições rígidas, as seguintes propriedades são válidas:

eq. 2-1 $$\bigcup_{i=1}^{c} a_i = X$$

eq. 2-2 $$a_i \cap a_j = f \quad \forall \text{ i} \neq \text{j}$$

eq. 2-3 $$f \subset a_i \subset X \quad \text{" i}$$

A

eq. 2-1 expressa o fato de que o espaço amostral é formado pela união de todas as categorias. A eq. 2-2 indica que nenhum elemento pode pertencer a mais de uma categoria e a eq. 2-3 mostra que nenhuma classe pode ficar vazia ou conter todos os elementos.

O valor de *c* deve variar entre $2 \leq c \leq n$, dado que para $c = 1$, todos os pontos pertencem à mesma categoria e para $c = n$, cada ponto pertence à sua própria categoria. Nestes dois casos, não existe de fato um problema categorização.

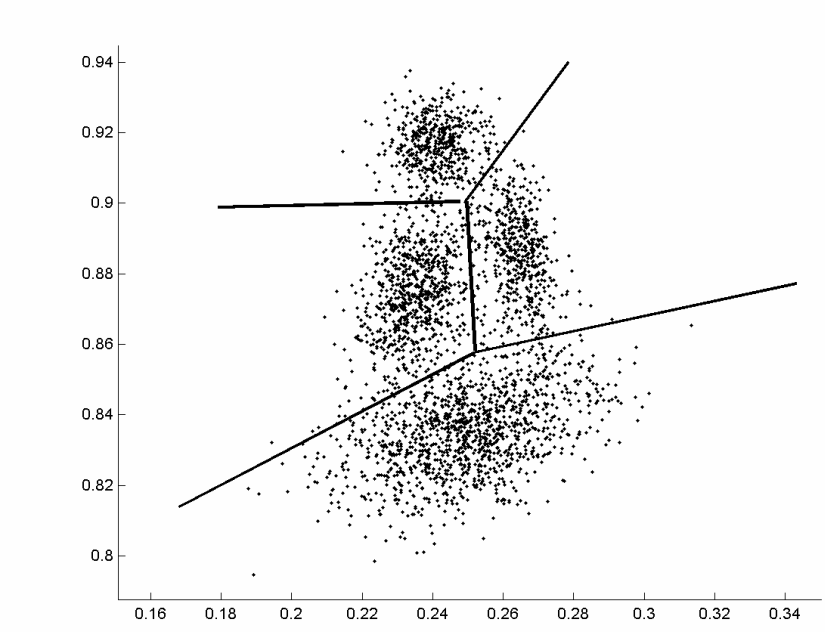

Figura 2.2 – Representação de 4 classes reais do $\hat{A}^2$ geradas aleatoriamente

Para o método *k-Means*, cada categoria é representada por um centro do conjunto $V = \{\mathbf{v_1}, ..., \mathbf{v_c}\}$. Neste método, cada amostra é agrupada na categoria que estiver mais próxima, ou seja, na categoria cuja *distância Euclidiana* da amostra ao centro é mínima.

A *distância Euclidiana* é dada pela fórmula:

eq. 2-4
$$d_{ik} = d(\mathbf{x_k}, \mathbf{v_i}) = \left\| \mathbf{x_k} - \mathbf{v_i} \right\| = \left[\sum_{j=1}^{p} \left(x_{kj} - v_{ij} \right)^2 \right]^{1/2}$$

A *matriz de partição* $U = \{u_{11}, ... , u_{cn}\}$ é a matriz das *funções características* u_{ij}, com *c* linhas e *n* colunas, que expressa a que categoria cada amostra pertence. O termo u_{ij} é a função característica do *j-ésimo* ponto na *i-ésima* categoria.

Os valores possíveis de u_{ij} são 1 se a amostra pertence à categoria ou 0 se a amostra não pertence à categoria. Suas propriedades são dadas por:

eq. 2-5
$$\sum_{i=1}^{c} u_{ij} = 1 \qquad \forall \text{ j}$$

eq. 2-6
$$0 < \sum_{j=1}^{n} u_{ij} < n \qquad \forall \text{ i}$$

eq. 2-7
$$u_{ij} = \begin{cases} 1 & u_{ij} \in \text{a}_\text{i} \\ 0 & u_{ii} \notin \text{a}_\text{i} \end{cases}$$

A eq. 2-5 juntamente com a eq. 2-7 garantem que uma amostra pertence a exatamente uma categoria e a eq. 2-6 indica que nenhuma categoria é vazia e que o número máximo de elementos é n-1.

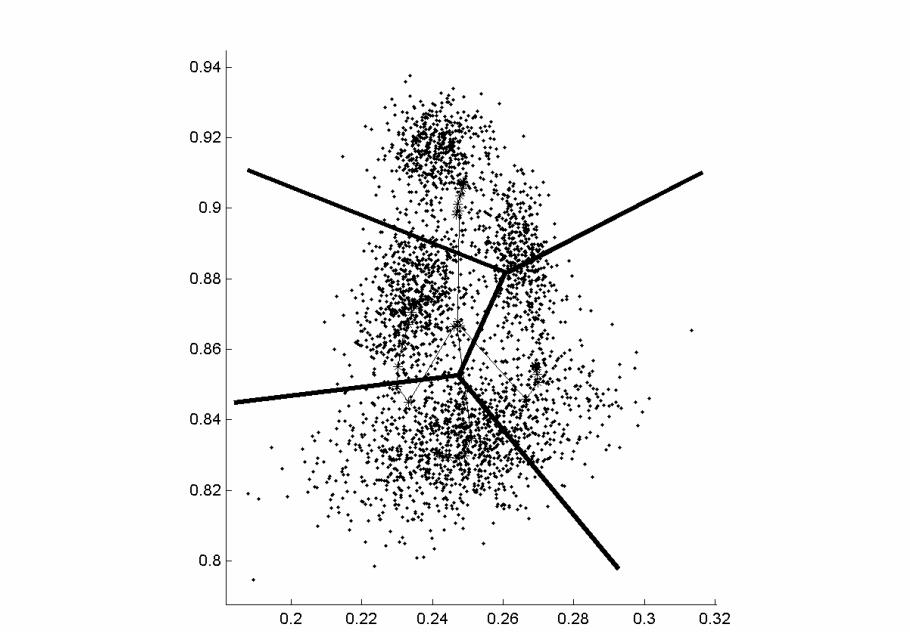

Figura 2.3 – Aplicação do método de categorização rígido k-Means a um problema com 4 classes do espaço $\hat{A}^2$. Pode-se observar o formato circular das categorias geradas.

O *k-Means* encontra a melhor partição rígida com *c* categorias para um espaço amostral minimizando a *função objetivo* J(U,V), que é dada pela fórmula:

$$J(U,V) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} (d_{ij})^2$$

No **Algoritmo k-Means**, é o *fator de parada* $\varepsilon > 0$ que determina o seu término. Quando a matriz *U* tem uma variação menor que ε entre duas iterações, o deslocamento dos centros das categorias no espaço é muito pequeno, não havendo uma mudança significativa no formato das partições. Isto indica que o algoritmo se estabilizou e uma solução eficiente foi encontrada.

Como o *k-Means* é um método rígido, o que implica que sua matriz *U* é rígida, o critério mais comumente utilizado é que a diferença entre a matriz *U* em duas iterações deve ser igual a 0, indicando que os centros não foram deslocados ou não tiveram um deslocamento significativo. Este critério foi empregado neste trabalho.

O *k-Means* é um método apropriado para ser aplicado quando a distribuição dos pontos das classes originais do problema pode ser aproximada por hiperesferas, sendo intrinsecamente dependente da disposição das amostras no espaço.

A Figura 2.2 ilustra 4 classes reais do $\Re^2$ para um conjunto de amostras geradas aleatoriamente. O método *k-Means* foi aplicado a este conjunto de amostras para 4 categorias e o resultado é mostrado na Figura 2.3.

Pode-se observar no resultado da categorização o formato esférico das categorias geradas e que estas não correspondem às classes do problema.

As deficiências de categorização do método podem ser analisadas a partir das categorias geradas. A categoria da parte superior da figura agrupou pontos das duas classes centrais, a categoria à direita absorveu pontos da classe inferior e a categoria da parte inferior da figura agrupou pontos da classe da esquerda.

Quanto mais próxima de hiperesferas é a distribuição dos dados, mais coesas são as categorias geradas. Para distribuições com outros formatos, as categorias hiperesféricas geradas contêm pontos de outras classes, comprometendo as categorias geradas pelo método.

Algoritmo k-Means:

Passo 1. Fixar o número de categorias nebulosas *c*, $2 \leq c < n$

Passo 2. Inicializar aleatoriamente *U*, obedecendo às equações **2-5** a **2-7**

Passo 4. Calcular o conjunto *V* dos *c* centros das categorias

$$\mathbf{v_i} = \frac{\sum_{k=1}^{n} u_{ik} \cdot \mathbf{x_k}}{\sum_{k=1}^{n} u_{ik}}$$

Passo 5. Recalcular a matriz *U* para os novos centros das categorias

se $d_{ij} = \min(d_{kj})$, $\forall\ k \in c$

/*se a categoria i é a mais próxima do ponto j */

então $u_{ij} = 1$

senão $u_{ij} = 0$

Passo 6. Comparar a nova matriz *U* com a anterior

se $\forall i,k\ \ u_{ik} - u_{ikanterior} \leq \varepsilon$ então fim

senão volte para o **Passo 4**

2.1.2 MÉTODO K-NN

O *K-NN* (K-Nearest Neighbour) é um método de classificação não supervisionado rígido. (KELLER,1985)

Um *método de classificação rígido* conhece a priori as classes existentes e associa cada amostra desconhecida a uma única classe. Cada classe é caracterizada por um conjunto de pontos rotulados denominados *padrões* e cada conjunto de padrões identifica uma e somente uma classe.

O número de padrões que representa cada classe pode ser variável, não sendo necessário que todas as classes tenham o mesmo número de padrões para defini-la.

O *K-NN* é um algoritmo que leva em conta a sua vizinhança. Seu objetivo é associar a amostra à classe que tiver mais padrões próximos a ela. Esta proximidade é calculada através da *distância Euclidiana*.

No caso de empate, ou seja, quando a amostra estiver próxima a um número igual de vizinhos para mais de uma classe, o critério de desempate é a soma das distâncias da amostra aos vizinhos de cada classe em que houve empate. O valor mínimo desta soma define a que classe a amostra pertence.

Se ainda ocorrer empate, a amostra pertencerá à classe cujo valor mínimo da soma for o último a ser calculado.

O método *K-NN* é um método interessante para ser aplicado quando as classes do problema são conhecidas e quando se quer associar cada amostra a uma única classe. Este método consegue aproximar qualquer distribuição de pontos, não se restringindo a uma distribuição específica das amostras.

Além disto, o método é muito útil quando existem poucas amostras de cada classe do problema, pois neste caso não é possível representar a distribuição probabilística dos dados que é necessária em outras técnicas, como os métodos GG e GK (seções 2.2.3 e 2.2.2)

Um dos problemas deste método rígido é que todos os padrões que representam as classes têm igual importância na classificação de uma amostra, o que dificulta a classificação quando há sobreposição alta das amostras de diferentes classes.

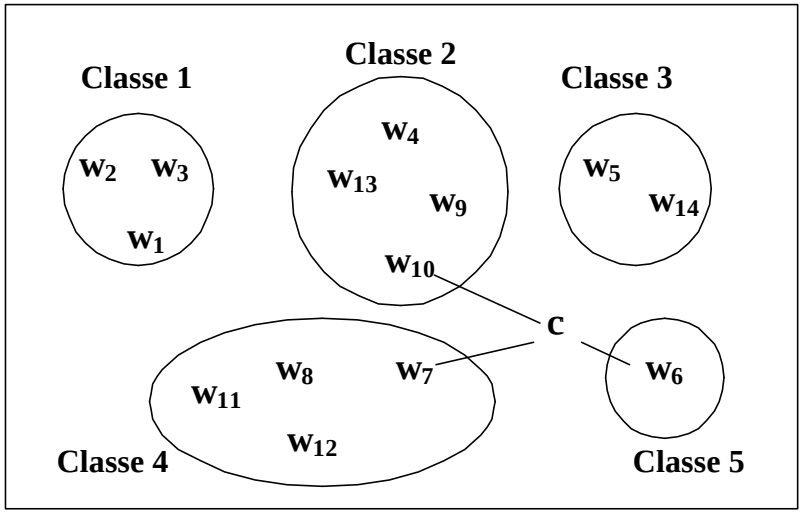

Figura 2.4 - K-NN rígido

Seja $W = \{\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_t}\}$ o conjunto de t amostras rotuladas do espaço $\Re^p$ que identificam as classes e χ a amostra a ser classificada, para o número de vizinhos $k = 3$, a Figura 2.4 ilustra cada classe sendo identificada pelos seus padrões e a classificação da amostra χ na classe 5 pela sua maior proximidade do padrão w_6.

Algoritmo K-NN rígido:

Passo 1. Fixar o valor de k, $1 \leq k \leq t$

Passo 2. **Para** $i = 1$ até t

Calcular a distância de χ aos padrões x_i

se $i \leq k$, incluir w_i no conjunto dos k vizinhos mais próximos

senão se w_i está mais próximo de χ que algum outro vizinho **então**

Apague o vizinho mais distante

Inclua w_i no conjunto dos k vizinhos mais próximos

Fim do se

Fim do Para

Passo 3. Determine a classe que é representada pelo maior número de padrões no conjunto dos k vizinhos mais próximos

Passo 4. **se** ocorrer um empate

Calcule a soma das distâncias de χ aos vizinhos em cada classe que empatou

se a soma for diferente

Classifique χ na classe cuja soma é mínima

senão

Classifique na classe onde o último mínimo foi encontrado

Fim do se

senão

Classifique na classe representada pelo maior número de padrões

Fim do se

2.2 MÉTODOS NEBULOSOS DE CATEGORIZAÇÃO E CLASSIFICAÇÃO

2.2.1 MÉTODO FUZZY C-MEANS

O *Fuzzy c-Means*, mais comumente chamado de *FCM*, é o método de categorização nebuloso mais utilizado por ser o mais simples de ser implementado, o mais rápido e por não infligir restrições ao conjunto de dados. (ROSS, 1997)

Ele é a versão nebulosa do método rígido *k-Means*, sendo empregado para classificar um universo de amostras em *categorias nebulosas* de acordo com a sua disposição no *Espaço Euclidiano*. Como no *k-Means*, as categorias geradas pelo *FCM* também são *hiperesferas* do $\Re^p$, de volume aproximado, caracterizadas pelos seus centros.

eq. 2-8
$$J_m(U,V) = \sum_{i=1}^{c}\sum_{j=1}^{n}\left(m_{ij}\right)^m \left(d_{ij}\right)^2$$

O *FCM* tem como finalidade minimizar a função objetivo J_m (eq. 2-8), sendo $\mathbf{x_i}$ um vetor com p características ou dimensões do conjunto não-rotulado $X = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}$ de n amostras, $V = \{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_c}\}$ o conjunto dos vetores que representam os c centros das categorias e U a matriz de graus de inclusão nebulosos, onde m_{ij} é o grau de inclusão do ponto j no cluster i.

A matriz *U* obedece as seguintes expressões:

eq. 2-9 $$\sum_{i=1}^{c} m_{ij} = 1 \qquad \forall \text{ j}$$

eq. 2-10 $$0 < \sum_{j=1}^{n} m_{ij} < n \qquad \forall \text{ i}$$

Na eq. 2-8, d_{ij} é a distância Euclidiana entre o centro do *i-ésimo* aglomerado e a *j-ésima* amostra. A eq. 2-9 indica que o somatório dos graus de inclusão de uma amostra em todas as categorias deve ser igual a 1 e a eq. 2-10 garante que nenhuma categoria pode ser vazia ou conter todos os elementos.

A constante nebulosa *m* é quem confere a nebulosidade às categorias resultantes. Ela pode assumir qualquer valor maior que 1. Quando $m \rightarrow 1$, a matriz *U* tende a ser rígida, sendo completamente rígida para $m = 1$.

Quando $m \rightarrow \infty$, os graus de inclusão tendem a $1/c$, ou seja, os pontos têm o mesmo grau de inclusão em todas as categorias. Não existe uma regra para definir o valor de *m* a ser usado, dependendo apenas de quão nebuloso ou rígido deseja-se que o particionamento do sistema seja. Os valores de *m* mais comumente usados são 1,25 e 2.

O número de categorias varia no intervalo $c \in [2 , n)$. O valor 1 é excluído do intervalo pois quando $c = 1$, todos os *n* pontos pertencem a uma única categoria.

O valor *n* também é excluído, pois quando $c = n$, cada centro coincide exatamente com um ponto do conjunto de amostras, tendo grau de inclusão 1 nesta categoria e zero nas outras. Deste modo, o método deixa de ser nebuloso funcionando precisamente como um método rígido.

Não existe uma norma para definir qual o número máximo de categorias ($c_{máx}$) a ser aplicado em um método de categorização. Segundo Pal e Bezdek (1995), "existe um pequeno direcionamento na literatura a respeito do $c_{máx}$. Uma regra simples que muitos investigadores utilizam é $c_{máx} \leq \sqrt{n}$".

Porém os autores desaconselham uma adesão estrita a esta regra, dado que em muitas situações existem alguns limites de $c_{máx}$ que são conhecidos pelo usuário.

Como no *k-Means*, o *fator de parada* ε é o valor que determina o término do algoritmo *FCM*, devido à estabilidade das partições ter sido alcançada. Como a matriz *U* é nebulosa, ε pode assumir qualquer valor, desde que mantenha a precisão desejada.

O uso do método *FCM* é vantajoso para categorizar dados não-rotulados cuja disposição no espaço pode ser aproximada por hiperesferas e no caso em que se sabe que as amostras têm características de mais de uma classe.

Outra vantagem do método é que ele não oferece restrições ao conjunto de dados, podendo ser aplicado a qualquer quantidade de amostras.

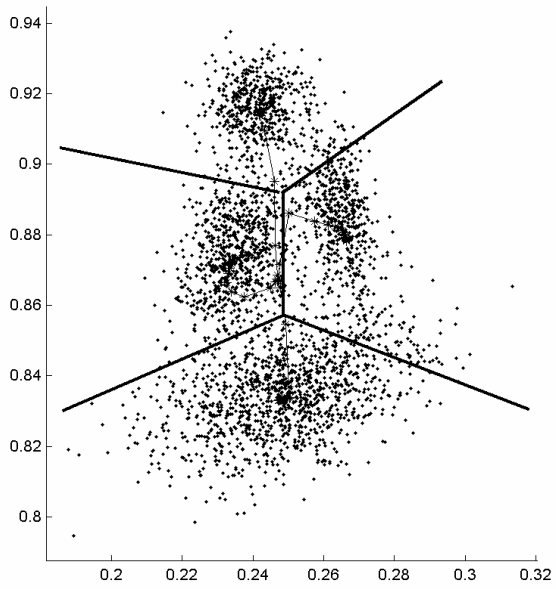

Figura 2.5 – Aplicação do método de categorização FCM a um problema com 4 classes do espaço $\hat{A}^2$. A categorização foi gerada usando m = 2 e c = 4. Pode-se observar o formato circular das categorias geradas.

A Figura 2.5 ilustra a aplicação do método *FCM* ao conjunto de amostras da Figura 2.2 para $c = 4$ e $m = 2$.

Para ilustrar o formato das categorias geradas, cada categoria foi delineada a partir das amostras com maior grau de inclusão nesta. Deste modo, temos um resultado rígido que não é necessariamente igual ao que seria obtido pelo método *k-Means*, dado que os métodos possuem características de convergência diferentes.

Pode-se observar claramente o formato circular das categorias geradas pelo *FCM* e que o resultado obtido é diferente do obtido pelo método *k-Means* (Figura 2.3), sendo que as categorias geradas pelo método *FCM* se aproximam mais do formato das classes do problema do que as geradas pelo *k-Means*.

Os erros de categorização podem ser observados a partir das categorias geradas. A categoria da parte superior da figura agrupou pontos das duas classes centrais originais, enquanto que a categoria da parte inferior também teve alguns de seus pontos capturados pelas categorias centrais da figura.

Algoritmo FCM:

Passo 1. Fixar o número de categorias nebulosos *c*, $2 \leq c < n$

Passo 2. Atribuir um valor à constante nebulosa *m* e ao fator de parada ε

Passo 3. Inicializar aleatoriamente a matriz *U*, obedecendo a normalização

$$\sum_{i=1}^{c} m_{ij} = 1$$

Passo 4. Calcular o conjunto *V* dos *c* centros das categorias

$$\mathbf{v_i} = \frac{\sum_{k=1}^{n} (m_{ik})^{m} \cdot \mathbf{x_k}}{\sum_{k=1}^{n} (m_{ik})^{m}}$$

Passo 5A. Calcular as distâncias de cada ponto a cada centro de aglomerado

$$d_{ik} = \left(\sum_{j=1}^{p} (x_{kj} - v_{ij})^2 \right)^{1/2}$$

Passo 5B. Recalcular a matriz *U* para os novos centros das categorias

se $d_{ik} > 0$, $\forall$ i $\in$ [1..c] **então** $m_{ik} = \left(\sum_{l=1}^{c} \left(\frac{d_{ik}}{d_{lk}} \right)^{\frac{2}{m-1}} \right)^{-1}$

senão /* Se alguma distância é zero */

se $d_{ik} = 0$ então $u_{ik} = 1$

senão $u_{ik} = 0$

Passo 6. Comparar a nova matriz *U* com a anterior
se $\forall i,k \; u_{ik} - u_{ikanterior} \leq \varepsilon$ então fim
senão volte para o **Passo 4**

2.2.2 MÉTODO GUSTAFSON-KESSEL

O método criado por Gustafson e Kessel, o método *GK*, é um método de categorização nebuloso, análogo ao método *FCM*, diferindo apenas na métrica da distância utilizada.

A distância Euclidiana empregada no *FCM* é substituída pela *Distância de Mahalanobis* (eq. 2-11) no **Passo 5A** do **Algoritmo FCM**. Neste caso, as categorias assumem a forma *hiperelipsóides* no espaço $\Re^p$ tendo aproximadamente o mesmo tamanho e o mesmo número de elementos. (GUSTAFSON, 1979)

O método *GK* oferece restrições ao conjunto de amostras utilizado, pois o cálculo da distância de Mahalanobis envolve a inversão da matriz de covariância fuzzy S_i (eq. 2-13).

Neste caso, para que S_i não seja singular, o número de amostras *n* deve ser maior que o número de dimensões dos dados mais 1, ou seja, maior que *p+1*.

Por outro lado, *p+1* amostras não fornecem uma boa estimativa de S_i. Como S_i contém p^2 elementos e é simétrica, ela contém $p(p-1)/2$ elementos independentes. Assim, estimativas válidas só são obtidas utilizando-se um número maior que $p(p-1)/2$ amostras. (VIVARELLI, 1999)

A *distância de Mahalanobis* é dada pela fórmula:

eq. 2-11
$$d_{ik}^{\,2} = (\mathbf{x}_k - \mathbf{v}_i)^T A_i (\mathbf{x}_k - \mathbf{v}_i)$$

onde as *matrizes de norma* A_i são dadas por

eq. 2-12

$$A_i = \sqrt[p]{\det(S_i)} \cdot S_i^{-1}$$

e a *matriz de covariância nebulosa* S_i é dada por

eq. 2-13

$$S_i = \frac{\sum_{j=1}^{n} m_{ij}^m (\mathbf{x_j} - \mathbf{v_i})(\mathbf{x_j} - \mathbf{v_i})^T}{\sum_{j=1}^{n} m_{ij}^m}$$

O método *GK* é um método que tem um bom desempenho quando a distribuição das amostras se aproxima de hiperelipses com o mesmo número de pontos.

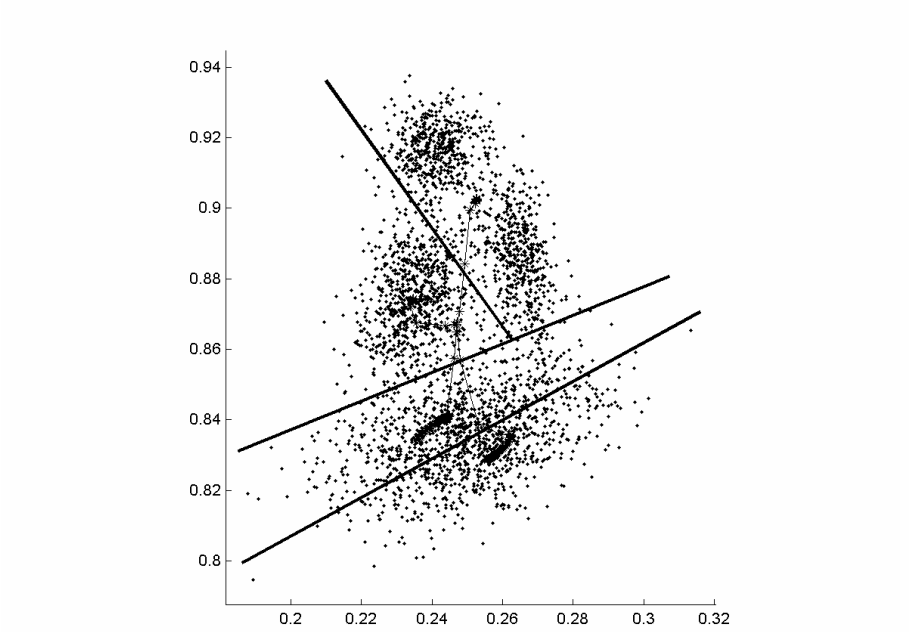

Figura 2.6 – Aplicação do método de categorização GK a um problema com 4 classes do espaço $\hat{A}^2$. A categorização foi gerada usando m = 2 e c = 4. Pode-se observar o formato elíptico das categorias geradas.

Quando os pontos não têm esta disposição ou quando os hiperelipsóides não têm o mesmo número de pontos, o método converge lentamente, gerando categorias que

pouco têm haver com as classes do problema. Mesmo quando obtém bons resultados, sua convergência é mais lenta que a do método *FCM*.

A Figura 2.6 ilustra a aplicação do método GK nos dados da Figura 2.2 para c = 4 e m = 2. Pode-se observar que as categorias assumem o formato elíptico.

As categorias geradas são bem diferentes das 4 classes reais do problema. A classe inferior foi basicamente dividida em duas categorias, enquanto que a classe da parte superior da figura foi agrupada com a classe da direita em uma única categoria.

Uma desvantagem do método é a imposição de restrições ao número de amostras, não podendo ser aplicado a amostras com muitas características quando a quantidade destas é pequena.

2.2.3 MÉTODO GATH-GEVA

Como o método *GK*, o método de Gath e Geva, conhecido como *método GG* ou *método de decomposição de mistura Gaussiana* (GMD - Gaussian Mixture Decomposition), também foi criado tendo como base o método *FCM*. (GATH, 1989)

A métrica de distância utilizada pelo método é a *Distância de Gauss* (eq. 2-14), que substitui a fórmula da distância Euclidiana no **Passo 5A** do **Algoritmo FCM**.

As categorias geradas pelo método *GG* já não têm mais formas geométricas definida, sendo estruturas no hiperespaço de tamanhos variados.

O método *GG* assume que os dados são uma mistura de c distribuições Gaussianas normais, sendo c o número de categorias. A *Distância de Gauss* é calculada a partir da *probabilidade a priori* P_i (eq. 2-16) de uma amostra pertencer a uma distribuição normal (categoria). P_i pode ser interpretado como um *parâmetro de tamanho da categoria*.

Assim as categorias que possuem mais elementos têm maior probabilidade de conter uma amostra que está sendo analisada, o que leva à tendência de uma categoria ser maior que outro.

Assim como o método *GK*, o *GG* também inflige restrições ao conjunto de amostras, pois a distância de Gauss também calcula a matriz de covariância nebulosa S_i (eq. 2-13).

A presença do termo exponencial no cálculo da distância de Gauss reduz a capacidade do *GG* de evitar mínimos locais, pois quando *U* é inicializado aleatoriamente, os centros gerados ficam geralmente muito próximos entre si e do centro global dos dados.

Para evitar este problema, é interessante inicializar o *GG* sempre pelos centros ao invés de *U* e certificar-se de que estes centros tenham sido obtidos por outros métodos de clusterização, como o *FCM*, e conseguir, assim, uma dusterização mais precisa.

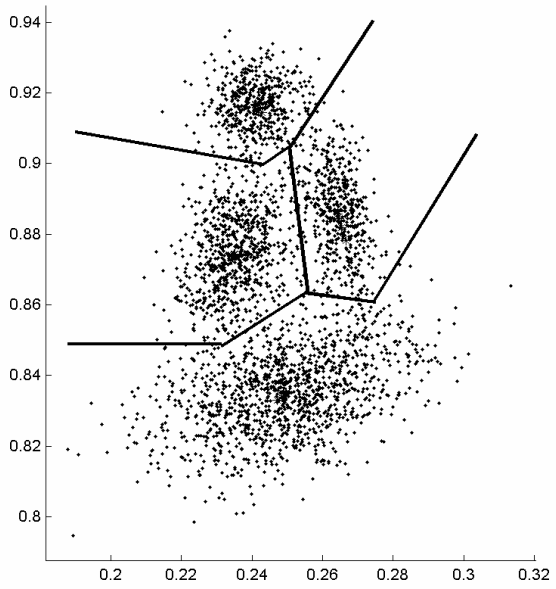

Figura 2.7 – Aplicação do método de categorização GG a um problema com 4 classes do espaço $\hat{\mathbf{A}}^2$. O método foi inicializado com os centros gerados pelo método FCM. Os parâmetros utilizados em ambos os algoritmos foram m = 2 e c = 4.

A *distância de Gauss* é dada pela fórmula:

eq. 2-14

$$d_{ik}{}^2 = \frac{\sqrt{\det(S_i)}}{P_i} \cdot \exp\left(\frac{1}{2}(\mathbf{x_k} - \mathbf{v_i})^T A_i (\mathbf{x_k} - \mathbf{v_i})\right)$$

onde as *matrizes de norma* A_i (norm matrices) tem a forma

eq. 2-15

$$A_i = S_i^{-1}$$

e são geradas à partir da *matriz de covariância nebulosa* S_i, (eq. 2-13).

A *probabilidade a priori* P_i é dada por

eq. 2-16

$$P_i = \frac{\sum_{j=1}^{n} m_{ij}^m}{\sum_{j=1}^{n}\sum_{t=1}^{c} m_{tj}^m}$$

A Figura 2.7 ilustra a categorização realizada pelo método GG nos dados mostrados na Figura 2.2. O método foi inicializado com os centros gerados pelo método *FCM*. Os dois algoritmos usaram os parâmetros $c = 4$ e $m = 2$.

O método consegue categorizar as 4 classes com bastante precisão, obtendo melhores resultados que o *FCM* e o *GK*. Pode-se observar que o formato indefinido das categorias foi o fator determinante da categorização correta do problema.

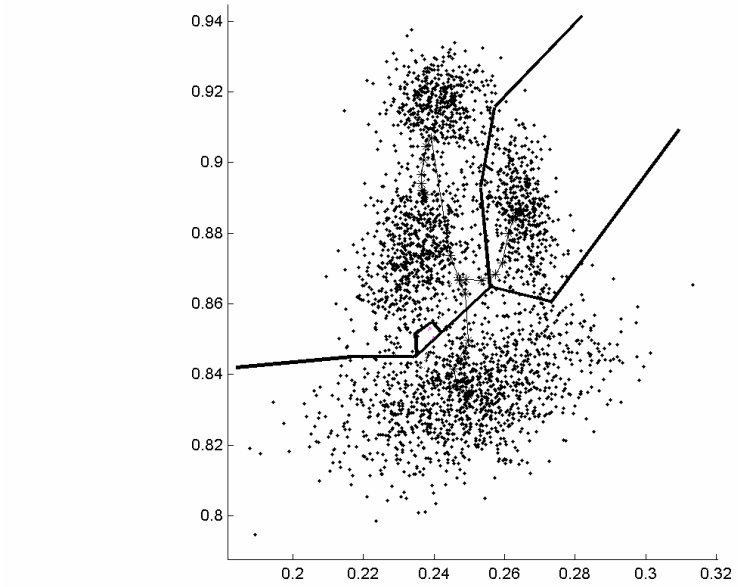

Figura 2.8 – Aplicação do método de categorização GG a um problema com 4 classes do espaço $\hat{A}^2$. O método gerou os centros sozinho. Pode-se observar o formato indefinido das categorias geradas

A Figura 2.8 ilustra a categorização feita pelo método por si só, com a inicialização aleatória da matriz U.

Neste caso, pode-se observar que o método categorizou erroneamente as classes do problema. A classe da direita e a inferior foram categorizadas quase corretamente, porém as classes superior e da esquerda foram assimiladas em uma única categoria enquanto uma quarta categoria com pouquíssimos pontos e que não equivale a nenhuma classe foi gerada.

O método *GG* é um método útil para aprimorar as categorias geradas pelo método *FCM*. Neste caso, o método é interessante porque consegue aproximar as categorias produzindo formas mais precisas que as hiperesferas obtidas pelo *FCM*.

Quando aplicado em conjunto com o *FCM*, ele converge rapidamente. Porém se aplicado sozinho, as categorias geradas pouco tem haver com as classes originais, pois ele tem uma grande tendência a convergir para mínimos locais. Neste caso, sua convergência é muito lenta.

Também se deve observar que o método *GG* também impõe restrições ao conjunto de amostras, não podendo ser aplicado em todos os casos.

2.2.4 MÉTODO FKCN

A *Rede Kohonen de Categorização Nebulosa FKCN* (Fuzzy Kohonen Clustering Network) é um método de categorização nebuloso não-supervisionado, derivado da *Rede Kohonen de Categorização KCN* (Kohonen Clustering Network). (GHOSH)

A *KCN* é uma rede neural fundamentada no *aprendizado competitivo - AC* (Competitive Learning).

Nesta técnica, apenas o neurônio cujo peso está mais próximo da amostra apresentada e um conjunto de neurônios conhecidos como seus vizinhos são atualizados para aprenderem a identificar a amostra. A métrica que estima a proximidade da amostra e dos neurônios é a *distância Euclidiana*.

O treinamento da rede *KCN* é *iterativo*, ou seja, os pesos da rede são atualizados após a apresentação de cada amostra. O conjunto de amostras é apresentado repetidas vezes à rede até que esta atinja a estabilidade.

Mas a *KCN* apresenta algumas desvantagens, como a falta de garantia da convergência e otimização do algoritmo de treinamento e a subutilização dos neurônios.

Devido ao seu treinamento *iterativo*, os critérios para atualização da *taxa de aprendizado* e da *vizinhança do neurônio vencedor*, assim como os *vetores de peso finais*, são dependentes das amostras e da ordem em que estas são apresentadas ao algoritmo.

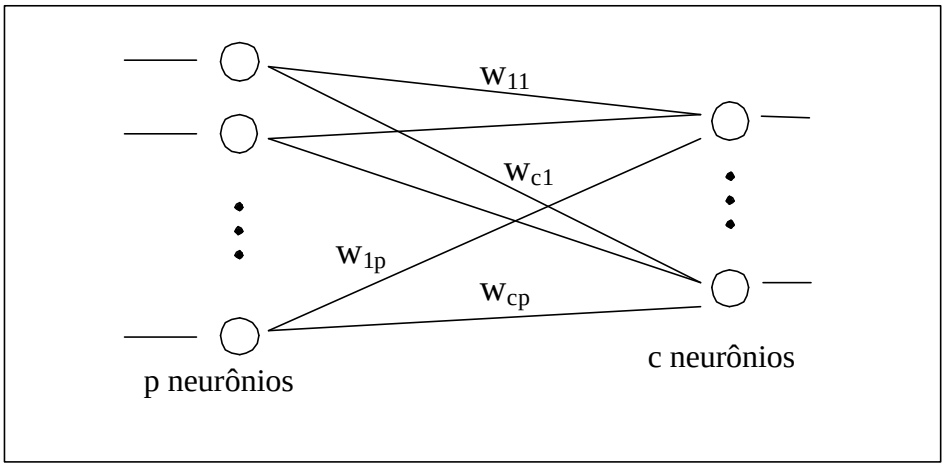

Figura 2.9 – Representação da Arquitetura da Rede Neural FKCN

Para solucionar estes problemas, a rede *FKCN* foi criada, baseada na integração dos métodos *FCM* e do *KCN*. Neste novo método, a taxa de aprendizado é controlada automaticamente, a topologia da vizinhança é dada pelos *graus de inclusão* e o treinamento é *batch*.

A arquitetura da *FKCN* (Figura 2.9) é formada por 2 camadas, uma de *entrada* e uma de *saída*. A *camada de entrada* é formada por *p* neurônios, onde *p* é o número de características das amostras. A *camada de saída* é formada por *c* neurônios, onde *c* é o número de categorias desejadas.

Seja $X = \{x_1, ... ,x_n\}$ um conjunto de *n* amostras *não-rotuladas* e $W = \{w_1, ..., w_c\}$ o conjunto dos vetores de peso dos neurônios de saída, ambos pertencentes ao espaço $\Re^p$.

Todos os *p* neurônios da camada de entrada estão conectados a todos os *c* neurônios da camada de saída com pesos individuais ajustáveis, onde w_{jk} é o peso da conexão entre o neurônio *j* da camada de saída e o neurônio *k* da camada de entrada.

Para cada vetor de treino x_i, os neurônios da camada de saída atualizam seus pesos baseados numa regra de aprendizado pré-definida.

A *taxa de aprendizado* α (eq. 2-21) e o critério de escolha dos neurônios vizinhos são reduzidos durante o treinamento a fim de alcançar a convergência da rede.

A constante nebulosa *m* também é reduzida durante as iterações até atingir *m* = 1 na iteração máxima. A cada iteração *m* é calculada pela fórmula:

eq. 2-17
$$m = m_0 - it \bullet \frac{m_0 - 1}{it_{\max}}$$

onde m_0 é a *constante nebulosa inicial* e deve ser inicializada com um valor maior que 1; *it* é o número da iteração atual e it_{max} é o *número máximo de iterações*.

O algoritmo termina quando o valor calculado *E* é menor que o fator de parada ε ou quando it_{max} é atingida. O valor de *E* é dado pela fórmula:

eq. 2-18
$$E = abs\left(\frac{D}{D_{ant}} - 1\right)$$

sendo D_{ant} o valor de *D* produzido na iteração anterior e *D* definido por

eq. 2-19
$$D = \frac{\sum_{i=1}^{n} \min_{j=1}^{c} \left\| x_i - w_j \right\|}{n}$$

O operador *min* da fórmula acima identifica a distancia Euclidiana mínima entre a amostra x_i e os pesos w_j dos neurônios da camada de saída.

Na primeira iteração, D_{ant} é calculado como os pesos gerados aleatoriamente no **Passo 1** do **Algoritmo FKCN**.

A rede *FKCN* gera categorias hiperesféricas e pode ser aplicada nos mesmos casos em que o método *FCM* é utilizado, não oferecendo restrições ao número de amostras quando estas têm um número de características grande.

Através das diversas execuções da rede realizadas para os dados da Figura 2.2, pôde-se observar que a rede *FKCN* converge mais rapidamente do que o método *FCM*, porém ela tem uma forte tendência a convergir para mínimos locais, gerando categorias que pouco representam as classes do problema.

Também se observou que ao convergir corretamente, a categorização gerada pela rede é quase idêntica à gerada pelo método *FCM*.

Algoritmo FKCN:

Passo 1. Inicializar aleatoriamente os vetores $w_j \in W$ e m

Passo 2. **Para** cada amostra $x_i \; \hat{I} \; X$, faça

Calcular o grau de inclusão μ_{ij} de x_i em cada uma das c categorias

eq. 2-20

$$m_{ij} = \left[\sum_{l=1}^{c} \left(\frac{\left\| \mathbf{x_i} - \mathbf{w_j} \right\|}{\left\| \mathbf{x_i} - \mathbf{w_l} \right\|} \right)^{\frac{1}{m-1}} \right]^{-1}$$

Calcular a taxa de aprendizado do vetor de peso w_j gerada pela amostra x_i

eq. 2-21

$$a_{ij} = \left(m_{ij} \right)^m$$

Atualizar os vetores de peso somando-os à Dw

eq. 2-22

$$\Delta w_j = \frac{\sum_{i=1}^{n} a_{ij} \left(\mathbf{x_i} - \mathbf{w_j} \right)}{\sum_{i=1}^{n} a_{ij}}$$

Fim do Para

Passo 4. se $it = it_{max}$ ou $E < \varepsilon$ terminar o algoritmo

senão voltar para o **Passo 2**

2.2.5 MÉTODO K-NN NEBULOSO

O *K-NN nebuloso* (fuzzy K-NN) é a versão nebulosa do algoritmo rígido *K-NN*. Ele é um método de classificação não supervisionado muito útil quando as categorias nas quais queremos associar cada amostra são conhecidas. (KELLER,1985)

Para o *K-NN nebuloso*, como no *K-NN rígido*, cada classe é caracterizada por um conjunto de pontos denominados *padrões* e cada classe pode ter um número de padrões diferente para representá-la.

Porém, no *K-NN nebuloso*, cada padrão tem um *grau de inclusão* em todas as classes existentes, ou seja, cada padrão pertence a todas as classes do problema.

Uma amostra é então classificada a partir dos *k* padrões mais próximos e dos seus respectivos *graus de inclusão* em cada uma das classes. Deste modo, cada amostra pode pertencer a uma ou mais classes com diferentes graus de inclusão, não existindo aqui o problema de empate.

Seja $W = \{\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_t}\}$ o conjunto dos *t* padrões que identificam as classes, *k* o número de vizinhos, m_{ij} o grau de inclusão do padrão $\mathbf{w_j}$ na classe *i* e $m_i(c)$ o grau de inclusão da amostra χ na classe *i*.

Na Figura 2.10, para $k = 3$, a amostra χ está mais próxima dos padrões w_6, w_7 e w_{10}. O padrão w_7 tem grau de inclusão diferente de zero nas classes 2 e 4. Com isto, χ tem um *grau de inclusão* diferente de zero nas classes 2, 4 e 5, pertencendo a estas classes, e igual a zero nas demais.

O *grau de inclusão* $m_i(c)$ da amostra nas classes $i = \{2,4,5\}$ é calculado a partir do *grau de inclusão* dos seus *k* vizinhos nas classes *i*, e é dado pela eq. 2-23.

O *K-NN nebuloso*, como o *K-NN rígido*, consegue aproximar qualquer distribuição de pontos, sendo um bom método a ser aplicado quando a distribuição dos pontos é desconhecida ou quando esta é diferente das distribuições impostas pelos outros métodos.

Como no *K-NN nebuloso* cada padrão tem um *grau de inclusão* em cada classe, o método não tem o mesmo problema que o *K-NN rígido* conseguindo classificar melhor as amostras quando a sobreposição destas é alta.

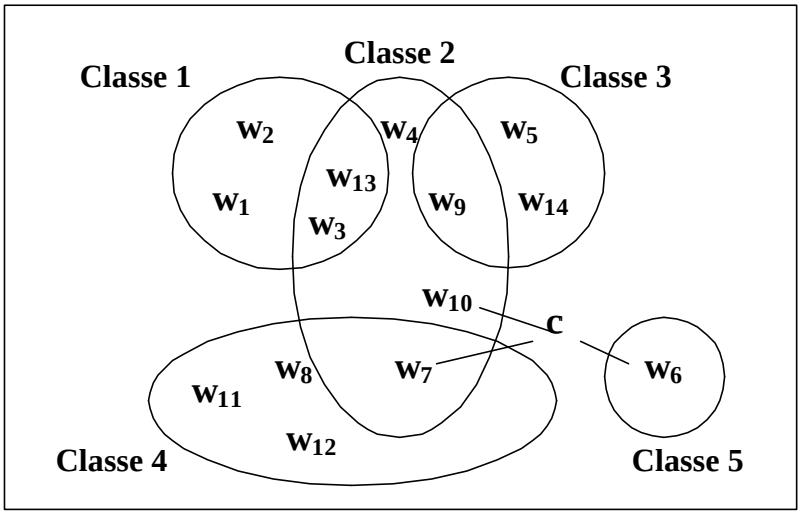

Figura 2.10 - K-NN nebuloso

Assim o método consegue associar cada amostra às classes com que tem maior afinidade, devendo ser utilizado quando é interessante saber o grau do relacionamento entre as amostras e as diversas classes do problema.

Algoritmo K-NN nebuloso:

Passo 1. Fixar o valor de k, $1 \leq k \leq t$

Passo 2. **Para** i = 1até t

Calcular a distância de χ aos padrões w_i

Se $i \leq k$, incluir x_i no conjunto dos k vizinhos mais próximos

Senão se x_i está mais próximo de χ que algum outro vizinho então

Apague o vizinho mais distante

Inclua x_i no conjunto dos k vizinhos mais próximos

Fim do Se

Fim do Para

Passo 3. Calcular o grau de inclusão $m_i(c)$ em todas as classes i

eq. 2-23

$$m_i(c) = \frac{\sum_{j=1}^{K} m_{ij} \left(1 / \left\| c - x_j \right\|^{2/m-1} \right)}{\sum_{j=1}^{K} \left(1 / \left\| c - x_j \right\|^{2/m-1} \right)}$$

2.3 MEDIDAS DE VALIDAÇÃO DE CATEGORIAS

Em alguns problemas de categorização, o número de classes ao qual queremos associar as amostras é previamente conhecido.

Um exemplo disto é a utilização do método de categorização nebuloso *FCM* para segmentação de imagens termais em sistemas de ar-condicionado. (WAKAMI, 1996)

O método *FCM* é aplicado para distinguir e separar o fundo de uma imagem térmica de uma sala refrigerada dos seus ocupantes. A imagem é representada por uma matriz de 8 x 20, onde seus elementos são a temperatura em cada ponto da sala.

Nesta aplicação, o número de classes é bem definido, onde o fundo da imagem representa uma classe e os ocupantes a outra. Um outro algoritmo é utilizado para calcular o número de ocupantes.

Mas existem casos em que ou as classes são desconhecidas ou o número de classes é conhecido mas não corresponde ao número de categorias que melhor representa o espaço amostral. Nestas situações é necessário utilizar uma métrica para identificar o melhor número de categorias a ser utilizado pelos métodos de categorização.

O estudo destas métricas é denominado *Validação de Categorias* (Cluster Validity) e as métricas geradas são chamadas de *medidas de validação* (validity measures).

As *medidas de validação* são aplicadas às partições de um conjunto amostral produzidas por um método de categorização a fim de estimar a qualidade dos particionamentos produzidos pelo método quando o número de categorias e as condições de inicialização são alterados.

Nesta seção serão apresentadas as *medidas de validação* de categorias nebulosas mais conhecidas e utilizadas na literatura. Também será apresentado o *Discriminante Linear de Fisher* – FLD (Fisher Linear Discriminant), que é um critério utilizado para validar categorias rígidas.

Estas medidas serão comparadas considerando suas funcionalidades e seus tempos de execução na seção 3.5.

2.3.1 COEFICIENTE DE PARTIÇÃO

A medida de validação *Coeficiente de Partição – F* (Partition Coefficient) indica o número ótimo de categorias de um espaço amostral quando seu valor máximo é atingido. (BEZDEK *apud* KOSANOVIC, 1995b)

Num espaço particionado, F assume valores no intervalo $1/c \leq F \leq 1$. Quando a matriz U é uma matriz rígida, ou seja, quando todos os graus de inclusão têm valores 1 ou 0, F assume o valor 1.

Para $F = 1/c$, o sistema atinge a maior nebulosidade possível, ou seja, cada ponto pertence a todas as categorias com o mesmo *grau de inclusão* ($1/c$), o que não é um bom resultado quando se almeja classificar amostras. F é dado pela fórmula:

eq. 2-24

$$F = \left(\sum_{i=1}^{c} \sum_{j=1}^{n} \left(m_{ij}\right)^2 \right) \Big/ n$$

A medida de validação F é estritamente nebulosa, ou seja, somente pode ser utilizada para analisar partições nebulosas.

Sua desvantagem é ser influenciada diretamente pelo número de categorias e pelo aumento da sobreposição das categorias, induzindo um comportamento decrescente à medida que o número de categorias aumenta.

Por este motivo, F não é a medida mais propícia para encontrar o número ideal de partições de um conjunto de dados, sendo mais apropriada para validar a melhor disposição das partições entre várias execuções de um método de categorização para um mesmo número de categorias.

2.3.2 *ENTROPIA DE PARTIÇÃO*

A medida de validação *Entropia de Partição – H* (Partition Entropy) também valida o número ideal de categorias para um espaço amostral particionado. Este número ideal é alcançado quando o valor mínimo de *H* é atingido. (BEZDEK *apud* KOSANOVIC, 1995b)

Os valores de *H* pertencem ao intervalo $0 \leq H \leq \log(c)$. Quando a matriz U é rígida, o valor de *H* é 0. Ele é dado pela fórmula:

eq. 2-25
$$H = -\left(\sum_{i=1}^{c} \sum_{j=1}^{n} m_{ij} \cdot \log\left(m_{ij}\right) \right) \Big/ n$$

Quando $H = \log(c)$, a nebulosidade máxima do sistema é atingida. Quando a matriz *U* é nebulosa, pode-se relacionar *H* e *F* pela desigualdade $0 \leq 1\text{-}F \leq H$.

A medida de validação *H* é estritamente nebulosa. Seu valor também é influenciado pelo número de categorias e pelo aumento da sobreposição destas, apresentando um comportamento crescente à medida que o número de categorias aumenta.

Deste modo, a medida *H* também é melhor empregada para validar a configuração das partições ótima de um conjunto amostral para as várias execuções de um método de categorização com um número específico de categorias.

2.3.3 *ÍNDICE DE PERFORMANCE DA NEBULOSIDADE*

O *Índice de Performance da Nebulosidade – FPI* (Fuzziness Performance Index) é uma medida de validação do número de categorias ideal de um conjunto amostral derivada do *coeficiente de partição – F*. (ROUBENS, 1982)

Ele estima o grau de nebulosidade de um sistema gerado por um método de categorização. O número ótimo de categorias é obtido pelo valor mínimo de *FPI*. Sua fórmula é dada por:

eq. 2-26

$$FPI = 1 - \frac{cF - 1}{c - 1}$$

Os valores de *FPI* pertencem ao intervalo $0 \leq FPI \leq 1$. Se a matriz U é rígida, o valor de *FPI* é 0, enquanto que para *FPI* = 1, o sistema atinge seu maior grau de nebulosidade.

A medida de validação *FPI* é estritamente nebulosa. Ao contrário de seu precursor, ele apresenta um comportamento crescente conforme a sobreposição das categorias aumenta, o que afeta sua confiabilidade.

2.3.4 ENTROPIA DE PARTIÇÃO MODIFICADA

A *Entropia de Partição Modificada – MPE* (Modified Partition Entropy) é uma medida de validação do número ideal de categorias baseada na *Entropia de Partição – H*. (ROUBENS, 1982)

Ela calcula o grau de desorganização gerado por cada número de categorias em que um espaço amostral foi dividido, sendo calculada pela fórmula:

eq. 2-27

$$MPE = \frac{H}{\log_a c}, (1 < \alpha < \infty)$$

Os valores de *MPE* estão no intervalo $0 \leq MPE \leq 1$. Quando *MPE*=0, a disposição das categorias do sistema se aproxima mais da forma rígida, enquanto que *MPE*=1 indica o maior grau de nebulosidade possível.

Quanto menor o valor de *MPE*, mais organizado é o sistema analisado e assim pode ser escolhido o melhor número de categorias deste.

A medida *MPE* também é estritamente nebulosa porém, ao contrário da medida *H*, não tem a restrição de aumentar conforme o número de categorias aumenta, sendo uma medida mais confiável que o *H*, mesmo quando a sobreposição das categorias é grande.

2.3.5 ÍNDICE NÃO NEBULOSO

O *Índice Não Nebuloso – NFI* (Nonfuzzy Index) é um índice que mede o número ideal de categorias do espaço amostral. Este número ótimo é obtido pelo valor máximo deste índice. (ROUBENS, 1978)

Os valores de *NFI* pertencem ao intervalo $0 \leq NFI \leq 1$. Quando a matriz U é rígida, o valor de *NFI* é 0 e quando $NFI = 1$, a nebulosidade máxima do sistema foi atingida. Ele é dado pela fórmula:

eq. 2-28

$$NFI = \left(c \cdot \left(\sum_{i=1}^{c} \sum_{j=1}^{n} \left(m_{ij} \right)^2 \right) - n \right) \Big/ n \cdot (c-1)$$

A medida *NFI* é estritamente nebulosa e apresenta a mesma desvantagem mostrada pelo *F*, perdendo precisão quando o número de categorias e a sobreposição destas aumentam.

2.3.6 TENDÊNCIAS RÍGIDAS MÍNIMA E MÉDIA

As *Tendências Rígidas Mínima e Média – MinHT e MeanHT* (Minimum and Mean Hard Tendencies) medem o quão bem definidas são as categorias quando consideradas as suas qualidades rígidas, sendo empregadas para calcular o número ótimo de categorias para um conjunto amostral. (RIVERA, 1990)

O *MeanHT* estima a média da tendência das categorias serem rígidas e o *MinHT* mede o mínimo da tendência que as categorias têm de serem rígidas. Eles assumem valores no intervalo $0 \leq MinHT\ ;\ MeanHT \leq \infty$ e quanto maiores os seus valores, mais definidas e compactas são as categorias analisadas.

O primeiro passo para calcular MinHT e o MeanHT é calcular a razão ri (

eq. 2-29) entre o segundo e o primeiro graus de inclusão máximos do ponto i (vetor Ui).

A razão r_i é dada pela fórmula:

eq. 2-29 $$r_i = \frac{m_{ki}}{m_{ji}}$$

onde $m_{ji} = \max_{1 \le t \le c} \{m_{ti}\}$ é o primeiro máximo de U_i

e $m_{ki} = \max_{\substack{1 \le t \le c \\ i \ne j}} \{m_{ti}\}$ é o segundo máximo de U_i

Como os *graus de inclusão* m_{ti} são normalizados, o valor de r_i é calculado no intervalo $0 \le r_i \le 1$.

Quando as partições são rígidas, o valor de r_i é zero para todos os pontos i = 1,2,...,n. Quando r_i=1, as partições atingem a nebulosidade máxima e são conseqüentemente indeterminadas.

O valor de r_i descreve, então, o quão rígido ($r_i \to 0$) ou o quão nebuloso ($r_i \to 1$) é um ponto x_i para uma determinada partição.

O próximo passo é transformar as partições nebulosas em partições rígidas (*desnebulizar*) associando cada ponto do espaço amostral a categoria em que ele tem o maior grau de inclusão. Esta associação é criada nos conjuntos Y_S.

eq. 2-30 $$Y_S = \left\{ x_i \,/\, m_{Si} = \max_{1 \le t \le c} \{m_{ti}\} \right\} \textbf{ com } s = 1,2,...,c$$

A seguir calcula-se a *tendência rígida* T_S (eq. 2-31) das categorias rígidas de Y_S. Ela é a média das razões r_i dos pontos x_i pertencentes à Y_S.

A *tendência rígida* T_S é dada pela fórmula:

eq. 2-31
$$T_S = \frac{\sum_{x_i \in Y_S} r_i}{card(Y_S)}$$

onde card(Y_S) é o número de pontos do conjunto Y_S, para card(Y_S) ≠ 0.

Os valores de T_S variam no intervalo $0 \leq T_S \leq 1$, com $T_S \rightarrow 0$ quando as partições nebulosas tendem às partições rígidas.

E, finalmente, os valores de *MinHT* e *MeanHT* são dados por:

eq. 2-32
$$MinHT = \max_{1 \leq s \leq c} \{-\log_{10}(T_S)\}$$

eq. 2-33
$$MeanHT = \frac{1}{c} \sum_{s=1}^{c} -\log_{10}(T_S)$$

A função $\log_{10}$ é introduzida nas equações acima tanto para aumentar a gama de valores possíveis das medidas como para torná-las positivas. Como dito acima, seus valores são computados no intervalo $0 \leq MinHT\,;\ MeanHT \leq \infty$.

O *MinHT* e o *MeanHT* são medidas estritamente nebulosas e são bastante confiáveis quando não há sobreposição das categorias, perdendo sua precisão com o aumento desta.

2.3.7 NEBULOSIDADES RELATIVAS MÍNIMA E MÁXIMA

As *Nebulosidades Relativas Mínima e Máxima – MinRF e MaxRF* (Minimum and Maximum Relative Fuzzyness) não são medidas de validação, mas sim *indicadores de validade*. (GORDON, 1992)

A função delas é calcular o grau de separação entre as categorias. Estas medidas são importantes pois quanto menor a sobreposição das partições, melhor categorizados estão os dados.

Quando *MinRF* e *MaxRF* assumem valores próximos de 1, a separação entre as categorias é ruim, enquanto que valores perto de zero sugerem uma boa separação entre as partições.

O *MinRF* e o *MaxRF* são calculados a partir dos conjuntos Y_S (eq. 2-30) definido na seção 2.3.6 para as medidas *MinHT* e *MeanHT*.

O primeiro passo é calcular a cardinalidade dos conjuntos Y_S, com card(Y_S) ≠ 0. O segundo valor a ser calculado é um *limiar L* (threshold) (eq. 2-34), que será utilizado para excluir pontos de Y_S com grau de inclusão maiores que ele. *L* é dado por:

eq. 2-34
$$L = {(1+c)}\big/{2c}$$

A *nebulosidade relativa – RF* (eq. 2-35) é a razão entre a cardinalidade de Y_S e o número de pontos de Y_S com grau de inclusão menor que *L* na categoria *S* ($m_{Si} < L$).

eq. 2-35
$$RF = \frac{card(Y_S)}{\underset{m_{Si}<L}{card}(Y_S)}$$

E, finalmente, *MinRF* e *MaxRF* são dadas por:

eq. 2-36
$$MinRF = \min(RF)$$

eq. 2-37
$$MaxRF = \max(RF)$$

O *MinRF* e o *MaxRF* são medidas estritamente nebulosas que conseguem avaliar uma configuração de categorias se estas são pouco sobrepostas. Se a sobreposição for alta, eles tendem erroneamente a indicar que a configuração com menos categorias tem o melhor grau de separação.

2.3.8 CARDINALIDADE MÍNIMA DE NMM

A *Cardinalidade Mínima da Aproximação Rígida por Grau de Inclusão Máximo – MinNMMcard* (Minimum Nearest Maximum Membership (NMM) Cardinality), assim como o *MinRF* e o *MaxRF*, também é um indicador de validade que estima a qualidade da separação entre as categorias. (KOSANOVIC, 1995a)

Os valores de *MinNMMcard* (eq. 2-38) variam de 0 a ∞. Quando o seu valor é igual a zero, significa que ou há categorias vazias ou o método de categorização nebuloso falhou ao tentar separar alguns dos centros das partições, ou seja, dois ou mais centros estão muito próximos de si.

O *MinNMMcard* também é calculado a partir dos conjuntos Y_S (eq. 2-30). Seu valor corresponde ao número de pontos da categoria desnebulizada que contém menos pontos. Sua fórmula é dada por:

eq. 2-38 $$MinNMMcard = \min\left(card\left(Y_S\right)\right)$$

O *MinNMMcard*, como o *MinRF* e o *MaxRF*, também avalia somente categorias nebulosas. Esta avaliação é confiável desde que as categorias estejam pouco sobrepostas, perdendo precisão quando a sobreposição é alta.

Os *indicadores de validade* têm maior funcionalidade quando associados a uma medida de validação. Por exemplo, para uma configuração de *c* categorias validadas, o *MinNMMcard* indica qual das execuções do método de categorização para as *c* categorias tem melhor grau de separação.

2.3.9 COMPACIDADE E SEPARAÇÃO

A *Compacidade e Separação – CS* (Compactness and Separation) é uma medida de validação mais completa, pois avalia tanto a compacidade das categorias geradas como a qualidade da separação entre estas. (XIE, 1991)

Quanto menor o valor de *CS* (eq. 2-39), melhor a disposição das categorias. Minimizar *CS* corresponde a minimizar a função objetivo J_m, que é a finalidade do método *FCM*.

Portanto a execução que obteve o menor valor de *CS* é a execução do método de categorização que foi mais bem sucedida na minimização da função objetivo, possuindo assim a melhor compacidade e separação. Para a *constante nebulosa m*, *CS* é escrita como:

eq. 2-39

$$CS = \frac{J_m}{n \cdot (d_{\min})^2}$$

A fórmula de *CS* pode ser alterada para validar categorias geradas por métodos de categorização com diferentes funções objetivo.

O fator d_{min} é a *distância Euclidiana* mínima entre dois centros de categorias. É ele quem mede a separação entre as categorias e é dado pela fórmula:

eq. 2-40

$$d_{\min} = \min_{i,j} \left\| \mathbf{v_i} - \mathbf{v_j} \right\|$$

Devido a sua implementação, a *CS* só é uma medida inócua quando o número de categorias tende a *n*. Mas isto não chega a ser um problema, pois *c* sempre é bem menor que *n* nos problemas de categorização.

Finalmente, expandindo a eq. 2-39, *CS* é dada pela fórmula:

eq. 2-41

$$CS = \frac{\sum_{i=1}^{c} \sum_{j=1}^{n} m_{ij}^2 \left\| \mathbf{v_i} - \mathbf{x_j} \right\|^2}{n \min_{i,j} \left\| \mathbf{v_i} - \mathbf{v_j} \right\|^2}$$

A medida *CS* é a mais completa e precisa medida de validação dentre as mostradas neste capítulo, pois além de validar o número de categorias, ela também

analisa o grau de separação entre elas, não perdendo exatidão mesmo quando a sobreposição das categorias é alta.

Além de avaliar partições nebulosas, a medida *CS* também pode ser usada para avaliar partições rígidas, desde que os graus de inclusão nebulosos sejam substituídos por graus de inclusão rígidos.

2.3.10 DISCRIMINANTE LINEAR DE FISHER

Na análise de problemas de classificação, quanto menor a dimensionalidade das características dos dados amostrais, mais fácil é a sua análise e a sua resolução sob o ponto de vista computacional.

O melhor espaço para analisar a disposição das amostras é o $\Re^1$ (Figura 2.11), quando estas são projetadas em uma reta, pois elas podem ser inspecionadas visualmente, evitando qualquer dúvida sobre a sua classificação.

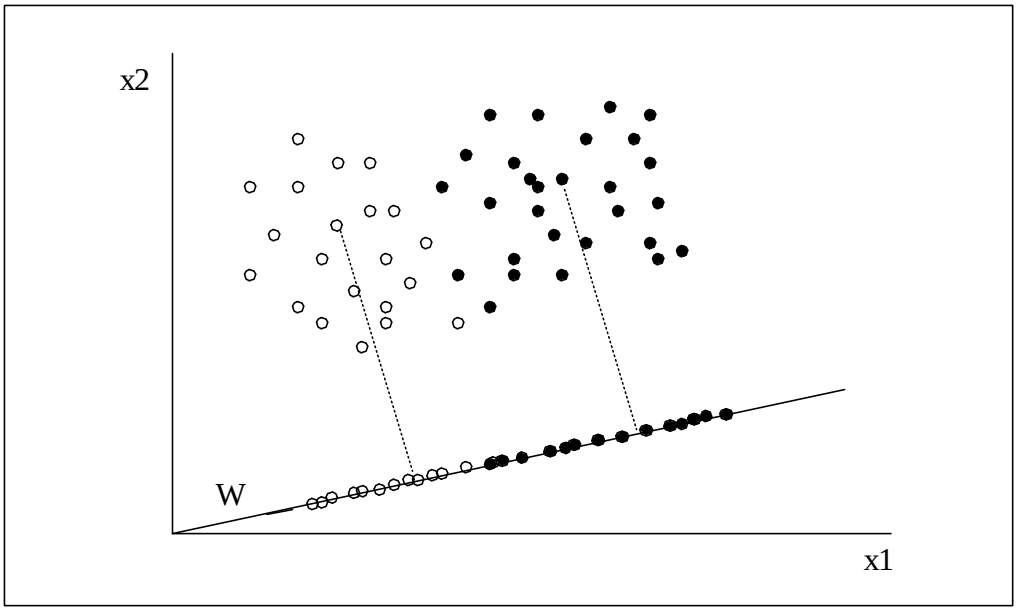

Figura 2.11 – Projeção de amostras dispostas em 2 classes em uma reta feita pelo Discriminante Linear de Fisher

Em vista disto, várias ferramentas foram criadas com o intuito de diminuir a dimensionalidade do espaço das características. Uma destas ferramentas é o *Discriminante Linear de Fisher – FLD* (Fisher Linear Discriminant). (BISHOP, 1995; DUDA, 1973)

O *FLD* é uma técnica supervisionada que, aplicada a um espaço rígido previamente particionado, obtém um operador de projeção linear *W* (eq. 2-49), que mapeia o conjunto de dados originais do $\Re^p$ para um novo espaço $\Re^k$ de menor dimensionalidade (k < p), com a propriedade de maximizar a separação entre as *c* categorias do espaço amostral.

A dimensão *k* do novo espaço amostral é uma função da distribuição das categorias das amostras. Seu valor pode estar entre 0 e *c* – 1, pois a maior base que *c* categorias produzem tem *c* – 1 dimensões.

A distribuição das categorias pode degenerar o tamanho da base produzida até chegar ao valor 0, que corresponde à situação em que todas as categorias possuem centros muito próximos.

O *FLD* mede a separação e a compacidade das categorias e do conjunto de todas as amostras através de *Matrizes de Espalhamento*.

A *Matriz de Espalhamento entre Classes* – S_B (Between-Class Scatter Matrix) mede a qualidade da separação entre as categorias. Para um conjunto de *n* amostras e *c* categorias, ela é dada por:

eq. 2-42
$$S_B = \sum_{i=1}^{c} n_i (\mathbf{m_i} - m)(\mathbf{m_i} - m)^T$$

onde *m* é a média do conjunto de todas as amostras e é dado pela fórmula

eq. 2-43
$$m = \frac{1}{n} \sum_{j=1}^{n} \mathbf{x_j}$$

e m_i é vetor média da classe c_i, onde n_i é o número de amostras da categoria *i*.

eq. 2-44
$$m_i = \frac{1}{n_i} \sum_{n \in c_i} x_i$$

A *Matriz de Espalhamento Interno* – S_W (Within-Class Scatter Matrix) mede a compacidade das categorias. Ela é a soma dos espalhamentos internos S_{Wi} de cada categoria c_i, dados pela equação:

eq. 2-45
$$S_{Wi} = \sum_{j \in c_i} (\mathbf{x_j} - \mathbf{m_i})(\mathbf{x_j} - \mathbf{m_i})^T$$

A fórmula final de S_W é:

eq. 2-46
$$S_W = \sum_{i=1}^{c} \sum_{j=1}^{n} (\mathbf{x_j} - \mathbf{m_i})(\mathbf{x_j} - \mathbf{m_i})^T$$

O *espalhamento total do sistema* – S_T é a soma do espalhamento interno das categorias e o espalhamento entre categorias, cuja fórmula é dada por:

eq. 2-47
$$S_T = S_W + S_B$$

Em um conjunto amostral com particionamento ótimo, a separação entre as categorias geradas deve ser máxima e o espalhamento dos dados de cada categoria deve ser mínimo. Neste caso, como o espalhamento total do sistema é independente das partições, observa-se que $S_B \rightarrow S_T$, forçando S_W a tender à matriz nula.

Para avaliar a qualidade da separação e da compacidade, Fisher define o *critério J*, que deve ser maximizado. Ele é a razão entre os determinantes de S_B e S_W e é dado pela fórmula:

eq. 2-48
$$J = \frac{|S_B|}{|S_W|}$$

Fisher utiliza o conceito de que o valor de *J* é afetado por operadores lineares de projeção para determinar o operador de projeção linear *W* que maximiza a função *J(W)*.

eq. 2-49

$$J(W) = \frac{\left|W^T S_B W\right|}{\left|W^T S_W W\right|}$$

O *critério J* (eq. 2-48) também pode ser usado como uma medida de validação de categorias rígidas, dado que quanto maior o seu valor, melhor a compacidade e separação do espaço amostral classificado.

Com este objetivo, um novo critério *J* (eq. 2-50) pode ser definido sem perda de precisão diminuindo, assim, o alto custo de computação imposto pelo cálculo dos determinantes das matrizes na fórmula acima.

O critério *J* é então calculado através da razão dos traços das matrizes de espalhamentos S_B e S_W, devendo ser *maximizado*. (BISHOP,1995)

J é então expresso por:

eq. 2-50

$$J = \frac{trace(S_B)}{trace(S_W)}$$

Os valores do critério *J* podem variar no intervalo $0 \leq J \leq \infty$. O *FLD* é uma medida estritamente rígida cuja desvantagem é a perda de precisão quando a sobreposição das amostras aumenta. Sua performance será avaliada na seção 3.5.

A tabela abaixo mostra o resumo das medidas de validação e de suas principais características que foram apresentadas nesta seção.

Medidas de Validação	Intervalo	Valor Ideal	Características	Partições
$F = \left(\sum_{i=1}^{c} \sum_{j=1}^{n} (m_{ij})^2 \right) \Big/ n$	$1/c \le F \le 1$	Maximizar	Afetada por c e pela sobreposição das amostras	N[1]
$H = -\left(\sum_{i=1}^{c} \sum_{j=1}^{n} m_{ij} \bullet \log(m_{ij}) \right) \Big/ n$	$0 \le H \le \log(c)$	Minimizar	Afetada por c e pela sobreposição das amostras	N
$FPI = 1 - \frac{cF - 1}{c - 1}$	$0 \le FPI \le 1$	Minimizar	Afetada pela sobreposição das amostras	N
$MPE = \frac{H}{\log_a c}, (1 < a < \infty)$	$0 \le MPE \le 1$	Minimizar	Não é afetada como sua precursora H	N
$NFI = \frac{\left(c \cdot \left(\sum_{i=1}^{c} \sum_{j=1}^{n} (m_{ij})^2 \right) - n \right)}{n \cdot (c-1)}$	$0 \le NFI \le 1$	Maximizar	Afetada por c e pela sobreposição das amostras	N
$MinHT = \max_{1 \le s \le c} \{ -\log_{10}(T_S) \}$	$0 \le MinHT \le \infty$	Maximizar	Afetada pela sobreposição das amostras	N
$MeanHT = \frac{1}{c} \sum_{s=1}^{c} -\log_{10}(T_S)$	$0 \le MeanHT \le \infty$	Maximizar	Afetada pela sobreposição das amostras	N
$MinRF = \min(\frac{card(Y_S)}{card(Y_S)})$ $m_{Si} < L$	$0 \le MinRF \le 1$	0 – BS[2] 1 – RS[3]	Indicador de validade que mede a qualidade da separação entre as categorias	N
$MaxRF = \max(\frac{card(Y_S)}{card(Y_S)})$ $m_{Si} < L$	$0 \le MaxRF \le 1$	0 – BS 1 – RS	Indicador de validade que mede a qualidade da separação entre as categorias	N
$MinNMMcard = \min(card(Y_S))$	$0 \le MinNMMcard \le \infty$	> 0 BS	Indicador de validade que mede a qualidade da separação entre as categorias	N

[1] N indica partições nebulosas
[2] BS significa Boa Separação
[3] RS significa Separação Ruim

Medidas de Validação	Intervalo	Valor Ideal	Características	Partições
$CS = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} m_{ij}^2 \left\|\mathbf{v_i} - \mathbf{x_j}\right\|^2}{n \min_{i,j} \left\|\mathbf{v_i} - \mathbf{v_j}\right\|^2}$	$0 \le CS \le \infty$	Minimizar	Ótima precisão. Não é afetada como as outras medidas	N/R
$FLD = J = \frac{trace(S_B)}{trace(S_W)}$	$0 \le FLD \le \infty$	Maximizar	Avalia a qualidade da compacidade e separação de partições rígidas	R[4]

Tabela 2.1 – Resumo das medidas de validação apresentadas nesta seção e de suas principais características

[4] R indica partições rígidas

3 PROPOSTA DE DUAS NOVAS MEDIDAS DE VALIDAÇÃO

Neste capítulo serão apresentadas as propostas de duas novas medidas de validação, o *Discriminante Linear de Fisher Estendido EFLD* (Extended Fisher Linear Discriminant) e a *Contraste entre Classes - ICC* (Inter Class Contrast). (FRANCO, 2002)

O *EFLD* é uma versão mais rápida da medida original, o *Discriminante Linear de Fisher – FLD*, que foi estendido para incluir a capacidade de avaliar tanto partições nebulosas como partições rígidas.

A *ICC* é a proposta de uma nova medida de validação rápida e eficiente, capaz de medir a compacidade e a separação de partições geradas por métodos de categorização rígidos e nebulosos sem sofrer a influência do número de categorias ou da sobreposição das amostras.

3.1 DESCRIÇÃO

O *Discriminante Linear de Fisher* pode ser utilizado como uma medida de validação de partições geradas por métodos de categorização rígidos.

As *técnicas de categorização rígidas* associam cada dado de entrada a uma e somente uma categoria, produzindo um mapeamento, de um para um, do conjunto de amostras para o conjunto de partições.

As *técnicas de categorização nebulosas*, ao contrário, consideram que cada ponto tem uma relação significativa com todas as categorias, onde o grau deste relacionamento é caracterizado pelo *grau de inclusão* m.

Para que o *FLD* pudesse manipular tanto partições nebulosas como rígidas, era necessário estendê-lo. Uma das propostas desta dissertação, mostrada no seção 3.2, é a

extensão do *FLD*, criando assim o *Discriminante Linear de Fisher Estendido EFLD* (Extended Fisher Linear Discriminant).

Após a criação e aplicação do *EFLD*, foi constatado que, assim como a versão tradicional, ele apresentava algumas falhas ao tentar validar categorias nebulosas e que estas falhas comprometiam o resultado das validações, como será mostrado nas seções 3.3 e 3.5.

A partir da análise destas deficiências, uma nova medida de validação é proposta na seção 3.4, chamada *Contraste entre Classes - ICC* (Inter Class Contrast).

Na seção 3.3 são mostrados os testes feitos com o *EFLD* e na seção 3.5 é mostrado o desempenho da medida de validação proposta *ICC* versus as medidas de validações apresentadas na seção 2.3.

3.2 PROPOSTA DA EXTENSÃO DO DISCRIMINANTE LINEAR DE FISHER

O *Discriminante Linear de Fisher Estendido EFLD* (Extended Fisher Linear Discriminant) é proposto neste capítulo como a versão nebulosa do *FLD*.

Considere um conjunto de dados não rotulados de n amostras $X = \{x_1, \ldots, x_n\}$ onde cada amostra x_i é um vetor com p características, $V = \{v_1, v_2, \ldots, v_c\}$ é o conjunto de centros das c categorias e U é a matriz de graus de inclusão, onde m_{ij} é o grau de inclusão do ponto j no cluster i.

Definição 1: S_{Be} é a matriz estendida de espalhamento entre classes, derivada da matriz S_B (eq. 2-42) rígida

eq. 3-1

$$S_{Be} = \sum_{i=1}^{c} \sum_{j=1}^{n} m_{ij} (\mathbf{m_{ei}} - \mathbf{m})(\mathbf{m_{ei}} - \mathbf{m})^T$$

onde m é o centróide do conjunto de todas as amostras

eq. 3-2

$$\mathbf{m} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{x_j}$$

$\mathbf{m_{ei}}$ é o centróide da categoria *i*

eq. 3-3

$$m_{ei} = \frac{\sum_{j=1}^{n} m_{ij}\mathbf{x_j}}{\grave{\imath}_{\mathbf{i}}}$$

e $\mathbf{m_i}$ é o somatório dos graus de inclusão de todas as amostras na categoria *i*

eq. 3-4

$$\grave{\imath}_{\mathbf{i}} = \sum_{j=1}^{n} m_{ij}$$

Definição 2: S_{We} é a matriz estendida de espalhamento interno, derivada da matriz S_W (eq. 2-46) rígida

eq. 3-5

$$S_{We} = \sum_{i=1}^{c}\sum_{j=1}^{n} m_{ij}\left(\mathbf{x_j} - \mathbf{m_{ei}}\right)\left(\mathbf{x_j} - \mathbf{m_{ei}}\right)^T$$

Definição 3: S_{Te} é a matriz estendida de espalhamento total, derivada da fórmula de S_T (eq. 2-47) rígida

eq. 3-6

$$S_{Te} = S_{We} + S_{Be}$$

Os passos que se seguem têm o objetivo de desenvolver a fórmula de S_{Te} definida na eq. 3-6, através da soma das matrizes de espalhamentos estendidas definidas nas equações eq. 3-1 e eq. 3-5.

Da eq. 3-3, tem-se que

eq. 3-7

$$\mathbf{m_{ei}} \sum_{j=1}^{n} m_{ij} = \sum_{j=1}^{n} m_{ij} \mathbf{x_j}$$

Expandindo S_{Be} (eq. 3-1)

$$S_{Be} = \sum_{i=1}^{c} \left[\left(\sum_{j=1}^{n} m_{ij} \right) \left(\mathbf{m_{ei}} \mathbf{m_{ei}^T} - \mathbf{m_{ei}} \mathbf{m^T} - \mathbf{m} \mathbf{m_{ei}^T} + \mathbf{m} \mathbf{m^T} \right) \right]$$

Desenvolvendo as parcelas de S_{Be}, tem-se

$$S_{Be} = \sum_{i=1}^{c} \left[\left(\sum_{j=1}^{n} m_{ij} \right) \mathbf{m_{ei}} \mathbf{m_{ei}^T} - \left(\sum_{j=1}^{n} m_{ij} \right) \mathbf{m_{ei}} \mathbf{m^T} - \left(\sum_{j=1}^{n} m_{ij} \right) \mathbf{m} \mathbf{m_{ei}^T} + \left(\sum_{j=1}^{n} m_{ij} \right) \mathbf{m} \mathbf{m^T} \right]$$

Aplicando a eq. 3-7 no segundo e no terceiro termos, tem-se a forma final de S_{Be}:

$$S_{Be} = \sum_{i=1}^{c} \left[\left(\sum_{j=1}^{n} m_{ij} \right) \mathbf{m_{ei}} \mathbf{m_{ei}^T} - \left(\sum_{j=1}^{n} m_{ij} \mathbf{x_j} \right) \mathbf{m^T} - \mathbf{m} \left(\sum_{j=1}^{n} m_{ij} \mathbf{x_j} \right)^T + \left(\sum_{j=1}^{n} m_{ij} \right) \mathbf{m} \mathbf{m^T} \right]$$

Expandindo S_{We} (eq. 3-5), obtém-se

$$S_{We} = \sum_{i=1}^{c} \sum_{j=1}^{n} m_{ij} \left(\mathbf{x_j} \mathbf{x_j^T} - \mathbf{x_j} \mathbf{m_{ei}^T} - \mathbf{m_{ei}} \mathbf{x_j^T} + \mathbf{m_{ei}} \mathbf{m_{ei}^T} \right)$$

Desenvolvendo as parcelas de S_{We}, tem-se

$$S_{We} = \sum_{i=1}^{c} \left[\sum_{j=1}^{n} m_{ij} \mathbf{x_j} \mathbf{x_j^T} - \left(\sum_{j=1}^{n} m_{ij} \mathbf{x_j} \right) \mathbf{m_{ei}^T} - \mathbf{m_{ei}} \left(\sum_{j=1}^{n} m_{ij} \mathbf{x_j} \right)^T + \left(\sum_{j=1}^{n} m_{ij} \right) \mathbf{m_{ei}} \mathbf{m_{ei}^T} \right]$$

Aplicando a eq. 3-7 no segundo e no terceiro termos, tem-se:

$$S_{We} = \sum_{i=1}^{c}\left[\sum_{j=1}^{n} m_{ij}\mathbf{x_j}\mathbf{x_j^T} - \left(\sum_{j=1}^{n} m_{ij}\right)\mathbf{m_{ei}}\mathbf{m_{ei}^T} - \left(\sum_{j=1}^{n} m_{ij}\right)\mathbf{m_{ei}}\mathbf{m_{ei}^T} + \left(\sum_{j=1}^{n} m_{ij}\right)\mathbf{m_{ei}}\mathbf{m_{ei}^T}\right]$$

E finalmente, S_{We} tem a forma

$$S_{We} = \sum_{i=1}^{c}\left[\sum_{j=1}^{n} m_{ij}\mathbf{x_j}\mathbf{x_j^T} - \left(\sum_{j=1}^{n} m_{ij}\right)\mathbf{m_{ei}}\mathbf{m_{ei}^T}\right]$$

Adicionando S_{Be} a S_{We}, temos

$$S_{We} + S_{Be} = \sum_{i=1}^{c}\left[\sum_{j=1}^{n} m_{ij}\mathbf{x_j}\mathbf{x_j^T} - \left(\sum_{j=1}^{n} m_{ij}\mathbf{x_j}\right)\mathbf{m^T} - \mathbf{m}\left(\sum_{j=1}^{n} m_{ij}\mathbf{x_j}\right) + \left(\sum_{j=1}^{n} m_{ij}\right)\mathbf{m}\mathbf{m^T}\right]$$

E, colocando os somatórios em evidencia,

$$S_{We} + S_{Be} = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\left(\mathbf{x_j}\mathbf{x_j^T} - \mathbf{x_j}\mathbf{m^T} - \mathbf{m}\mathbf{x_j^T} + \mathbf{m}\mathbf{m^T}\right)$$

Reagrupando, temos:

eq. 3-8
$$S_{We} + S_{Be} = \sum_{j=1}^{n}\left(\sum_{i=1}^{c} u_{ij}\right)\left(\mathbf{x_j} - \mathbf{m}\right)\left(\mathbf{x_j} - \mathbf{m}\right)^T = S_{Te}$$

A eq. 3-8 expressa a equação da matriz de espalhamento total estendida S_{Te}. Pode-se observar que, *quando o somatório dos graus de inclusão de uma amostra em todos as categorias é definido como 1* (eq. 2-9*), a matriz nebulosa de espalhamento*

total S_{Te} é igual a matriz rígida de espalhamento total S_T definida por Fisher, como ilustra a eq. 3-9.

Com este resultado provou-se que o espalhamento total de um sistema é independente da natureza das partições, sejam estas nebulosas ou rígidas.

eq. 3-9
$$\forall j, \sum_{i=1}^{c} m_{ij} = 1 \quad \Rightarrow \quad S_{We} + S_{Be} = \sum_{j=1}^{n} (\mathbf{x_j} - \mathbf{m})(\mathbf{x_j} - \mathbf{m})^T = S_T$$

Assim é definido neste trabalho o critério J_e que é a extensão do critério J mostrado na eq. 2-48, sendo calculado a partir dos determinantes das matrizes de espalhamento estendidas S_{Be} (eq. 3-1) e S_{We} (eq. 3-5). Seu valor deve ser maximizado a fim de validar a qualidade da separação e da compacidade das categorias analisadas. Sua fórmula é dada por:

eq. 3-10
$$J_e = \frac{|S_{Be}|}{|S_{We}|}$$

J_e também pode ser expresso pela razão dos traços das matrizes de espalhamentos S_{Be} (eq. 3-1) e S_{We} (eq. 3-5), sendo a extensão de J definido na eq. 2-50, e sendo expresso por:

eq. 3-11
$$J_e = \frac{trace(S_{Be})}{trace(S_{We})}$$

A eq. 3-11 é uma boa forma de avaliar partições geradas por métodos de categorização, dado que o traço das matrizes é mais rápido de calcular do que seus respectivos determinantes e não impõe limites em relação ao número mínimo de pontos em cada partição.

O cálculo deste critério pode ser ainda mais otimizado observando-se que as matrizes de espalhamento são geradas pelo produto de um vetor coluna pelo seu

transposto e portanto seus traços correspondem ao quadrado do módulo de seu vetor gerador.

Deste modo, os traços das equações eq. 3-1 e eq. 3-5 na eq. 3-11 podem ser reescritos pelos os escalares s_{Be} e s_{We}:

eq. 3-12
$$s_{Be} = trace(S_{Be}) = \sum_{i=1}^{c}\sum_{j=1}^{n} m_{ij}\left\|\mathbf{m_{ei}} - \mathbf{m}\right\|^2$$

eq. 3-13
$$s_{We} = trace(S_{We}) = \sum_{i=1}^{c}\sum_{j=1}^{n} m_{ij}\left\|\mathbf{x_j} - \mathbf{m_{ei}}\right\|^2$$

A otimização do critério J_e continua partir da observação de que a soma dos traços das matrizes S_{We} e S_{Be} é constante para o conjunto de amostras analisado. O traço de matriz S_T (s_T) é dado pela equação:

eq. 3-14
$$s_T = trace(S_T) = \sum_{j=1}^{n}\left\|\mathbf{x_j} - \mathbf{m}\right\|^2$$

Assim, o *EFLD* pode ser reescrito como a equação eq. 3-15, que é mais rápida de calcular, dado que o termo s_T precisa ser calculado apenas uma vez para o conjunto de amostras e que o termo s_{Be} é mais rápido de calcular do que s_{We}. As comparações de tempo de execução das diversas formas do critério de Fisher serão mostradas na seção 3.5.

eq. 3-15
$$J_e = \frac{s_{Be}}{s_T - s_{Be}}$$

3.3 APLICANDO O EFLD

O critério J_e do *EFLD*, definido na eq. 3-15, pode ser utilizado como uma medida de validação para o número ideal de categorias geradas por métodos de categorização nebulosos.

O valor máximo de J_e indica o melhor número de categorias para um sistema nebuloso. Porém os valores do critério J_e, como acontece com os gerados pelo critério J (eq. 2-50) do *FLD*, tem a tendência a crescer conforme o número de categorias e a sobreposição das amostras aumentam.

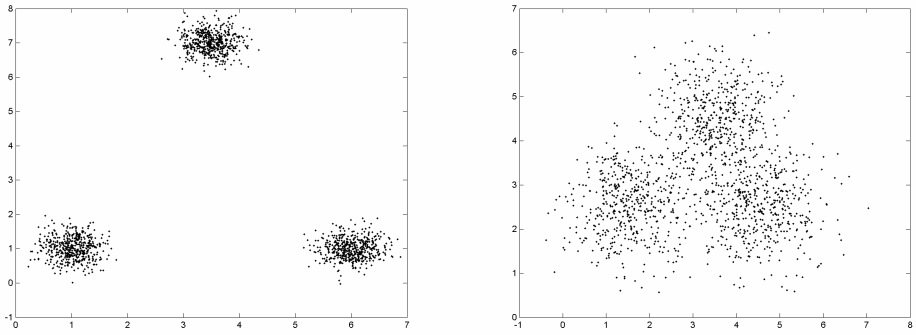

Figura 3.1 – Amostras X1 – 3 agrupamentos sem sobreposição e Amostras X2 – 3 agrupamentos com sobreposição alta

Para avaliar o critério J_e foram geradas aleatoriamente duas amostras de dados, contendo cada uma três classes de 500 pontos cada, com distribuição normal. A Figura 3.1 mostra a disposição destas amostras.

As três classes em que estão dispostos os pontos do conjunto de amostras *X1* estão bem separadas umas das outras, não havendo sobreposição destas. Suas classes foram centradas nos pontos (1,1), (6,1) e (3,5, 7), com desvio padrão 0,3 para os dois eixos.

As amostras do conjunto *X2* estão distribuídas por três classes com alta sobreposição de seus pontos. Suas classes foram centradas nos pontos (1,5, 2,5), (4,5, 2,5) e (3,5, 4,5) com desvio padrão 0,7 em ambos os eixos.

O algoritmo de categorização nebuloso *FCM* foi aplicado a ambos os conjuntos para $m = 2$ e com o número de categorias c variando de 2 a 6. Em seguida, o EFLD foi

aplicado a todas as partições geradas a fim de avaliar qual o melhor número de categorias para cada conjunto de amostras.

Pela Tabela 3.1, observa-se que para as amostras *X1* sem sobreposição, o critério de *EFLD* acerta o número de categorias como sendo 3.

Porém, para as amostras *X2* que estão muito sobrepostas, o critério de *EFLD* mostra a sua tendência a crescer conforme o número de categorias aumenta, indicando como 6 o número ideal de partições nebulosas.

EFLD	**Número de Categorias**				
	2	**3**	**4**	**5**	**6**
Amostras X1	4,6815	**4,9136**	0,2943	0,2559	0,3157
Amostras X2	0,3271	0,8589	0,8757	0,9608	**1,0674**

Tabela 3.1 - Valor do critério de EFLD para as amostras X1 e X2, após a execução do FCM com o número de agrupamentos variando de 2 a 6 categorias para m = 2

Por este motivo, o critério J_e do *EFLD* não é uma medida confiável para descobrir o melhor número de categorias, principalmente se a sobreposição das amostras do sistema que está sendo analisado é alta.

Porém o critério do *EFLD* tem uma função importante para os algoritmos de categorização.

Se o algoritmo de categorização cair num mínimo local não conseguindo distribuir os centros pelo sistema de amostras, deixando-os todos juntos, o valor numérico gerado pelo critério do *EFLD* é extremamente pequeno e sabe-se que o algoritmo de categorização deve ser executado novamente.

A Figura 3.2 mostra a aplicação do algoritmo de categorização *FCM* no conjunto de amostras *X1* para 2 categorias e $m = 2$. O algoritmo não conseguiu distribuir os dois centros no sistema, representados pelos dois pontos no centro da figura, caindo num mínimo local que coincidiu com o ponto médio do conjunto de todos os pontos.

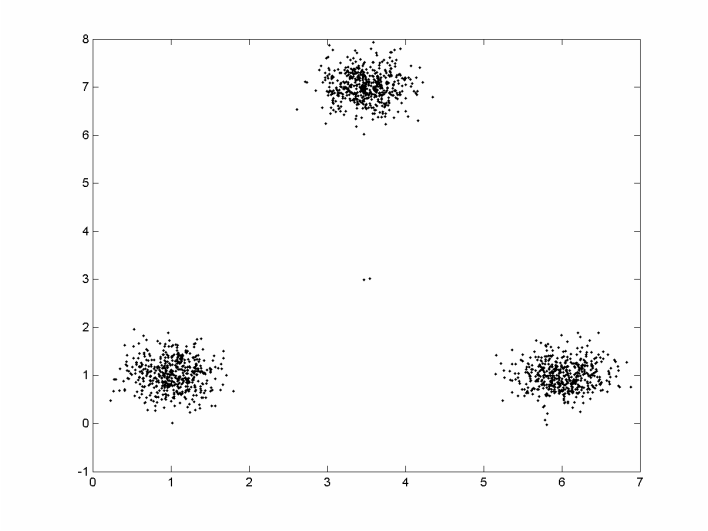

Figura 3.2 - Centros gerados pelo FCM para a amostra X1 com c =2 e m = 2. Os 2 centros caíram em um mínimo local coincidente com o ponto médio do conjunto de dados

O valor do critério do *EFLD* calculado para esta situação é 9,8010 x 10^{-5}, indicando a má disposição dos centros e a necessidade de executar novamente o algoritmo de categorização para estes centros.

3.4 CONTRASTE ENTRE CLASSES: MEDIDA DE VALIDAÇÃO PROPOSTA

Nesta seção está sendo proposta a medida de validação *Contraste Entre Classes – ICC* (Inter Class Contrast), criada empiricamente a partir da observação feita sobre a funcionalidade e as restrições do critério de *EFLD*, descritas na seção acima.

O *EFLD* tem a tendência de crescer conforme o número de partições e a sobreposição das amostras aumentam, atingindo seu valor máximo para um falso número ideal de categorias em um conjunto de amostras analisado.

Este problema ocorre quando os algoritmos de categorização são obrigados a inserir mais de um centro em cada classe do problema, decorrente do fato de que o número de categorias é maior que o número de classes reais.

A *ICC* foi criada para ser capaz de avaliar um espaço particionado por uma ferramenta de categorização nebulosa ou rígida, levando em conta a separação entre as categorias geradas e a compacidade destas.

Ela também foi moldada para detectar centros alocados muito próximos, o que compromete uma boa categorização. Quanto maior o seu valor, melhor o particionamento do conjunto de dados. Sua fórmula é dada por:

eq. 3-16

$$ICC = \frac{s_{Be}}{n} \bullet D_{\min} \bullet \sqrt{c}$$

O s_{Be}, definido na eq. 3-12, é o termo do *EFLD* que estima a qualidade da alocação do centros das categorias, onde centros mal alocados produzem valores baixos de s_{Be}.

Para evitar o falso crescimento da *ICC* quando o número de categorias é maior que o número de classes, devido ao comportamento do termo s_{Be}, o termo D_{min} foi acrescentado à sua fórmula, dado que quando duas ou mais categorias são associadas a uma mesma classe, a distância mínima entre os centros D_{min} decresce abruptamente.

Assim, D_{min} evita que o máximo valor de *ICC* seja atingido para um valor de *c* maior que o valor ideal. D_{min} é a distância Euclidiana mínima entre os centros das categorias e é dado pela fórmula:

eq. 3-17

$$D_{\min} = \min_{1 \le i \le c} \left[\min_{i+1 \le j \le c-1} \left\| m_{ei} - m_{ej} \right\| \right]$$

Quando uma ou mais categorias englobam mais de uma classe, ou seja, quando o número de categorias é menor que o número de classes, a distância mínima D_{min} entre os centros aumenta, aumentando o valor da *ICC*.

Para evitar que a *ICC* atinja seu valor máximo para um valor de *c* menor que o ótimo, a raiz quadrada do número de categorias foi introduzida em sua fórmula. Esta condição pode ocorrer quando uma ou mais categorias representam mais de uma classe do problema e os centros destas categorias estão longe um dos outros, gerando valores altos de D_{min} e conseqüentemente valores altos da medida *ICC*.

Assim, o termo $\sqrt{c}$ garante que a medida *ICC* cresça juntamente com o número de categorias, alcançando seus valores máximos próximos do valor ótimo de c, enquanto D_{min} evita que o valor máximo de *ICC* seja atingido para um valor de *c* maior que o valor ótimo.

O fator 1/n é um fator de escala, usado para compensar a influência do número de pontos no termo s_{Be}.

3.5 APLICANDO A ICC

Esta seção mostra a análise do desempenho da medida de validação proposta *ICC* ao validar partições nebulosas e rígidas geradas pelos algoritmos *FCM* e *k-Means*, respectivamente em relação às medidas de validação descritas na seção 2.3, ao *FLD* e à medida *EFLD* que também foi proposta neste trabalho.

A Tabela 3.2. apresenta todas as medidas de validação que serão testadas e comparadas com a *ICC*. As fórmulas das medidas de validações nebulosas encontram-se na Tabela 2.1.

Para mostrar a diferença do tempo de cálculo da *ICC* com o uso do determinante e do traço em relação ao uso do escalar s_{fe}, as medidas *ICCDet* e *ICCTra* são definidas. Elas usam a fórmula da medida *ICC* (eq. 3-16), trocando-se o escalar s_{Be} pelo determinante e o traço da matriz S_{Be}, respectivamente.

A medida *EFLD* é o critério J_e, definido pela eq. 3-15. As medidas *EFLDTra* e *EFLDDet* são as funções critério dadas pelas fórmulas eq. 3-10 e eq. 3-11, respectivamente.

Os testes foram executados utilizando 2 conjuntos de amostras *X1* e *X2* ilustrados na Figura 3.3. Cada conjunto foi gerado com cinco classes randômicas bidimensionais com distribuição normal de 500 pontos cada.

As classes do primeiro conjunto de amostras *X1* foram centradas nos pontos (1, 2), (6, 2), (1, 6), (6, 6) e (3,5, 9), com desvio padrão de 0.3 para os dois eixos (Figura 3.3).

ICC	Maximizar (M)
ICCDet	Maximizar (M)
ICCTra	Maximizar (M)
EFLD	Maximizar (M)
EFLDTra	Maximizar (M)
EFLDDet	Maximizar (M)
FLD	Maximizar (M)
CS	Minimizar (m)
NFI	Maximizar (M)
F	Maximizar (M)
FPI	Minimizar (m)
H	Minimizar (m)
MPE	Minimizar (m)
MinHT	Maximizar (M)
MeanHT	Maximizar (M)
MinRF	próximo a 0 – separação boa (0) próximo a 1 – separação ruim
MaxRF	próximo a 0 – separação boa (0) próximo a 1 – separação ruim
MinNMMcard	próximo a 0 – centros mal alocados (>0)

Tabela 3.2 - Resumo dos melhores casos para as Medidas de Validação. *M* indica que a medida deve ser maximizada, *m* indica que a medida deve ser minimizada, (0) indica que a medida deve estar próxima de 0 e (>0) o valor da medida deve ser maior que 0.

As classes do conjunto de dados *X2* foram centradas nos pontos (2, 2,5), (4, 2,5), (3, 7), (2, 5) e (4, 5), com desvio padrão de 0.7 para os dois eixos (Figura 3.4).

O método de categorização nebuloso *FCM* foi aplicado aos dois conjuntos para a constante nebulosa $m = 2$, variando o número de categorias de 2 a 10. Em seguida, cada um dos critérios de validação foi aplicado às categorias geradas a fim de determinar o melhor número de partições e os tempos de execução.

Para o conjunto *X1* (Figura 3.3) formado por 5 classes bem separadas, com superposição nula, observa-se na Tabela 3.3 que todas as medidas de validação, exceto a *MinHT* concordam que o melhor número de categorias é 5.

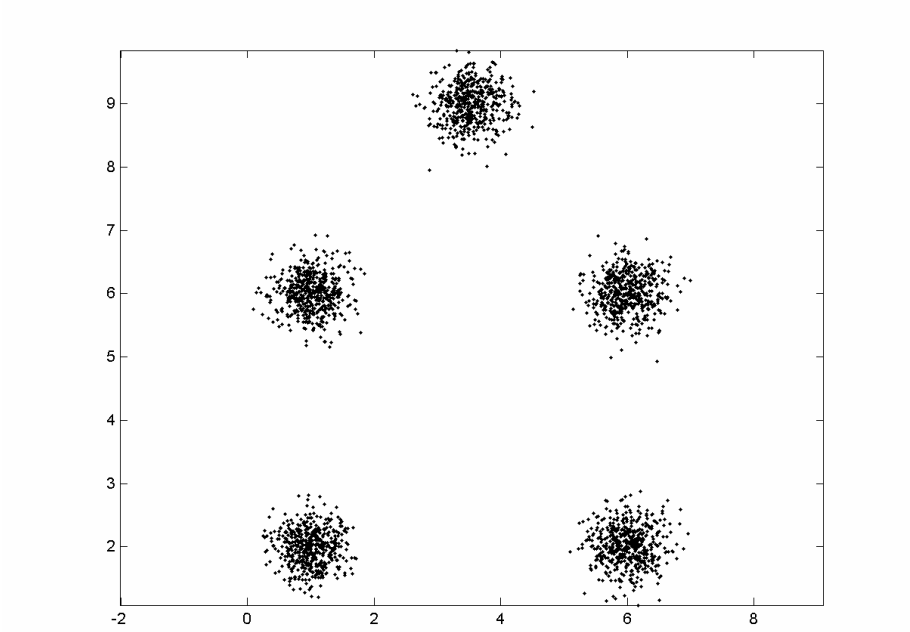

Figura 3.3 - Conjunto das amostras X1 com 5 agrupamentos centrados nos pontos (1;2), (6;2), (1;6), (6;6) e (3.5;9) e desvio padrão de 0.3 para os dois eixos

A medida de validação *MinHT* valida 4 como o número de categorias ótimo para *X1* ao invés de 5. Também se pode observar que para 9 categorias, os indicadores de validade *MinRF*, *MaxRF* e *MinNMMcard* desqualificam a categorização avaliada, indicando uma má alocação dos centros das categorias.

Medidas		**Número de Categorias**								
		2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
ICC	**M**	7,596	41,99	51,92	**96,70**	8,728	8,076	8,772	0,0002	10,86
ICCTra	**M**	7,596	41,99	51,92	**96,70**	8,728	8,076	8,772	0,0002	10,865
ICCDet	**M**	IND	154685	259791	**673637**	60571	55923	60916	0,0012	74865
EFLD	**M**	0.185	0.986	1.877	**13.65**	12.95	12.73	12.92	0.0008	11.72
EFLDTra	**M**	0,185	0,986	1,877	**13,65**	12,95	12,73	12,92	0,0008	11,72
EFLDDet	**M**	IND	0,955	3,960	**182,70**	164,46	157,75	165,14	IND	135,88
CS	**m**	0,350	0,096	0,070	**0,011**	0,734	0,809	0,697	22285	0,584
NFI	**M**	0,411	0,569	0,727	**0,9294**	0,851	0,776	0,715	0,0002	0,686
F	**M**	0,705	0,713	0,795	**0,943**	0,876	0,808	0,751	0,111	0,717
FPI	**m**	0,588	0,430	0,272	**0,070**	0,148	0,223	0,284	0,999	0,313
H	**m**	0,452	0,509	0,413	**0,157**	0,269	0,377	0,470	2,196	0,574
MPE	**m**	0,653	0,463	0,298	**0,097**	0,150	0,194	0,226	0,999	0,249
MinHT	**M**	0,647	0,572	**2,124**	1,994	1,993	1,993	1,864	0,026	1,991
MeanHT	**M**	0,519	0,496	1,327	**1,887**	1,409	1,044	0,787	0,006	0,733
MinRF	**0**	0,100	0,316	0	**0**	0	0	0	1	0
MaxRF	**0**	0,589	0,476	0,442	**0**	0,167	0,169	0,152	1	0,428
MinNMMcard	**>0**	1059	731	500	**500**	237	246	242	0	96

Tabela 3.3 - Valores das Medidas de Validação para o conjunto das amostras X1 particionado pelo FCM para 2 a 10 categorias nebulosas e m = 2

Os valores de *ICC* e *ICCTra*, assim como os valores de *EFLD* e *EFLDTra* são idênticos, porém o tempo de execução (Tabela 3.4) de *ICC* é muito menor que o de *ICCTra*, e o tempo de execução de *EFDL* também é menor que o tempo de *EFDLTra*, mostrando serem métodos de validação melhores e mais rápidos.

Estas medidas são mais rápidas do que suas respectivas versões utilizando o determinante das matrizes de espalhamento.

Para *c=2*, as medidas *ICCDet* e *EFLDDet* são indeterminados (observe o valor IND na Tabela 3.3) para o conjunto de amostras *X1*, pois a matriz S_{Bf} é singular, sendo impossível calcular o seu determinante. Isto ocorre por causa da distribuição das categorias que produz centros simétricos em relação ao centro global dos pontos.

Os valores de *MinRF* indicam uma boa separação para um número de categorias variando de 4 a 8 agrupamentos. Já *MaxRF* indica que a melhor separação ocorre com exatamente 5 categorias.

A *MinNNNcard* valida como boa todas as separações dos centros efetuadas pelo algoritmo *FCM*, exceto para 9 categorias.

Tempos	**Número de Categorias**								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
ICC	0,0061	0,0069	0,0082	**0,0091**[5]	0,0090	0,0107	0,0113	0,0113	0,0124
ICCTra	0,0078	0,0060	0,0088	0,0110	0,0122	0,0110	0,0170	0,0154	0,0170
ICCDet	0,0110	0,0088	0,0110	0,0132	0,0160	0,0142	0,0186	0,0160	0,0174
EFLD	0,0053	0,0071	0,0063	**0,0080**[4]	0,0093	0,0113	0,0107	0,0118	0,0121
EFLDTra	0,7678	1,0870	1,4780	1,8982	2,2422	2,7310	3,0594	3,098	3,587
EFLDDet	0,7800	1,1392	1,5510	2,0160	2,5870	2,8010	3,5150	3,50	4,13
CS	0,0226	0,0261	0,0382	0,0476	0,0508	0,0645	0,0719	0,0738	0,0756
NFI	0,0061	0,0056	0,0058	**0,0060**[3]	0,0066	0,0067	0,0074	0,0063	0,0069
F	0,0044	0,0045	0,0049	**0,0049**[1]	0,0061	0,0074	0,0079	0,0151	0,0156
FPI	0,0061	0,0045	0,0049	**0,0053**[2]	0,0065	0,0084	0,0090	0,0135	0,0068
H	0,0184	0,0206	0,0259	0,0313	0,0374	0,0492	0,0593	0,0681	0,0890
MPE	0,0176	0,0236	0,0266	0,0327	0,0442	0,0611	0,0599	0,0731	0,0725
MinHT	0,0453	0,0469	0,0481	0,0505	0,0593	0,0530	0,0569	0,0582	0,0645
MeanHT	0,0412	0,0448	0,0475	0,0546	0,0558	0,0525	0,0571	0,0664	0,0626
MinRF	0,0423	0,0500	0,0596	0,0580	0,0549	0,0547	0,0675	0,0601	0,0695
MaxRF	0,0525	0,0678	0,0511	0,0571	0,0577	0,0660	0,0753	0,1198	0,1436
MinNMMcard	0,0088	0,0104	0,0132	0,0126	0,0138	0,0156	0,0173	0,0373	0,0418

Tabela 3.4 - Tempos de execução em segundos das Medidas de Validação para o conjunto das amostras X1 particionado pelo FCM para 2 a 10 categorias nebulosas e m = 2. Os índices 1 a 4 indicam a posição das medidas de validação em relação ao tempo de execução.

As medidas *F*, *NFI* e *FPI* e a medida proposta *EFLD* são as medidas de validação mais rápidas nesta ordem (ver Tabela 3.4), seguidas pela *ICC*. Todas as

medidas com determinante são mais lentas que as demais e pode-se notar que a *ICCTra*, que é calculada com o traço da matriz, é mais lenta que a *ICC* que usa o escalar s_{Bf} (eq. 3-16).

No conjunto de amostras *X2* (Figura 3.4), os pontos das classes estão bem espalhados e há uma alta sobreposição das classes.

Para este conjunto, as medidas *F*, *H*, *NFI*, *FPI*, *MinHT e MeanHt* validam 2 categorias (Tabela 3.5), que é o menor número de categorias testado, como o melhor número de categorias para o conjunto de dados.

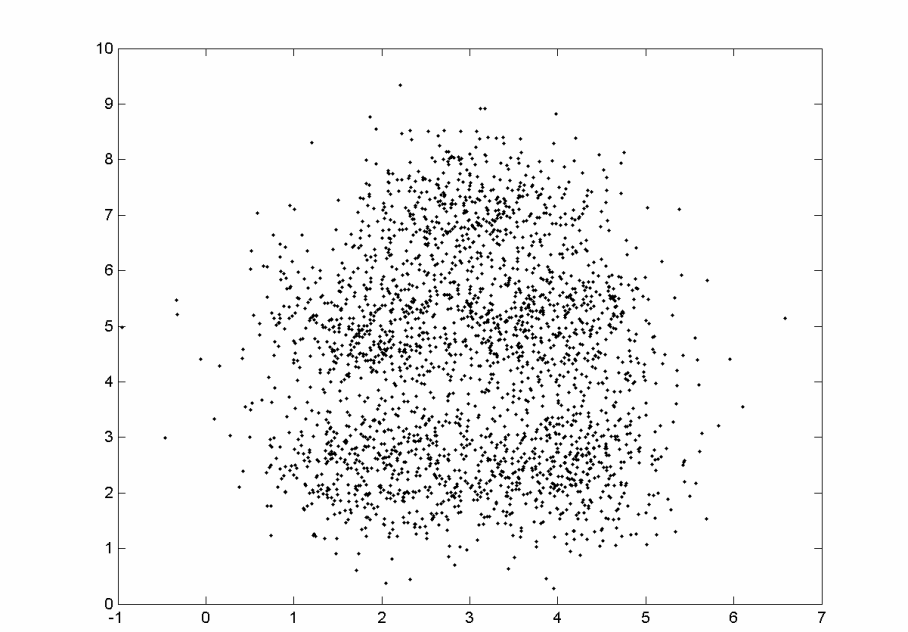

Figura 3.4 - Conjunto das amostras X2 com 5 agrupamentos centrados nos pontos (1;2), (6;2), (1;6), (6;6) e (3.5;9) e desvio padrão de 0.7 para os dois eixos

Também se pode observar que os valores de *H*, *EFLD* e *EFLDTra* são monotonamente crescentes quando o número de categorias cresce, enquanto que os valores de *F* decrescem proporcionalmente ao número de partições.

Os indicadores *MinRF* e *MaxRF* também classificam 2 categorias como a melhor disposição de categorias para *X2*. A *MinNMMcard* confirma uma boa separação para todos as partições geradas pelo *FCM*, tendo maior valor para 2 categorias.

As medidas *ICCDet* e *EFLDDet* também são indeterminados para a execução do FCM usando $c = 2$ e o conjunto de amostras *X2*.

Medidas		Número de Categorias								
		2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
ICC	**M**	5,065	4,938	6,191	**7,829**	6,837	6,121	6,130	6,21	5,69
ICCTra	**M**	5,065	4,938	6,191	**7,829**	6,837	6,121	6,130	6,21	5,69
ICCDet	**M**	IND	715,19	3572	**7048**	6178	5870	6136	6390	6024
EFLD	**M**	0.450	0.585	0.839	1.095	1.100	1.181	1.237	1.291	**1.344**
EFLDTra	**M**	0,450	0,585	0,839	1,095	1,100	1,181	1,237	1,291	**1,344**
EFLDDet	**M**	IND	0,049	0,315	0,743	0,748	0,887	1,004	1,096	**1,200**
CS	**m**	0,164	0,225	0,191	**0,122**	0,218	0,217	0,225	0,217	0,223
NFI	**M**	**0,508**	0,432	0,455	0,482	0,437	0,412	0,399	0,385	0,376
F	**M**	**0,754**	0,621	0,591	0,586	0,530	0,496	0,474	0,453	0,439
FPI	**m**	**0,491**	0,567	0,544	0,517	0,562	0,587	0,600	0,614	0,623
H	**m**	**0,393**	0,660	0,777	0,846	0,991	1,087	1,167	1,239	1,301
MPE	**m**	0,568	0,601	0,561	**0,525**	0,553	0,559	0,561	0,564	0,565
MinHT	**M**	**0,670**	0,536	0,600	0,654	0,603	0,554	0,520	0,506	0,478
MeanHT	**M**	**0,632**	0,485	0,550	0,597	0,529	0,474	0,463	0,438	0,429
MinRF	**0**	**0,170**	0,294	0,194	0,210	0,248	0,320	0,334	0,373	0,402
MaxRF	**0**	**0,280**	0,498	0,548	0,356	0,533	0,518	0,534	0,539	0,559
MinNMMcard	**>0**	1160	720	523	472	302	274	250	239	217

Tabela 3.5 - Valores das Medidas de Validação para o conjunto das amostras X2 particionado pelo FCM para 2 a 10 categorias nebulosas e m = 2.

As únicas medidas que indicaram o número de categorias ótimo como 5 foram *CS*, *MPE* e *ICC* com suas variações. Em termos de tempo de execução, a medida de validação *ICC* é mais rápida que as outras duas, ficando em primeiro lugar, seguida da *MPE* e da *CS* nesta ordem.

Tempos	Número de Categorias								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
ICC	0,0060	0,0064	0,0077	**0,0088**[1]	0,0093	0,0099	0,0110	0,0121	0,0148
ICCTra	0,0066	0,0060	0,0098	0,0110	0,0122	0,0110	0,0152	0,0170	0,0198
ICCDet	0,0110	0,0078	0,0110	0,0120	0,0160	0,0132	0,0170	0,0230	0,0220
EFLD	0.0063	0.0088	0.0096	0.0110	0.0096	0.0113	0.0116	0.0126	0.0127
EFLDTra	0,7930	2,1038	1,7598	2,2584	2,6784	3,0142	3,4922	3,3560	3,7070
EFLDDet	0,9720	1,2580	1,6090	1,8450	2,2470	2,6090	3,5920	3,3450	3,6580
CS	0,0220	0,0283	0,0362	**0,0590**[3]	0,0645	0,0728	0,0879	0,1494	0,0919
NFI	0,0044	0,0049	0,0058	0,0057	0,0063	0,0068	0,0102	0,0077	0,0074
F	0,0112	0,0121	0,0061	0,0164	0,0143	0,0066	0,0066	0,0074	0,0080
FPI	0,0052	0,0047	0,0072	0,0076	0,0090	0,0069	0,0069	0,0076	0,0099
H	0,0362	0,0351	0,0428	0,0690	0,1055	0,0802	0,0709	0,0695	0,0683
MPE	0,0167	0,0271	0,0319	**0,0397**[2]	0,0599	0,0491	0,0557	0,0621	0,0689
MinHT	0,0453	0,0508	0,1294	0,0783	0,0678	0,0676	0,0692	0,0632	0,0651
MeanHT	0,0458	0,0478	0,0544	0,0563	0,0829	0,0903	0,0860	0,0810	0,0977
MinRF	0,0549	0,0472	0,0476	0,0516	0,0549	0,0571	0,0687	0,0739	0,0909
MaxRF	0,0507	0,0612	0,0558	0,0582	0,0634	0,0632	0,0788	0,0775	0,0931
MinNMMcard	0,0089	0,0107	0,0121	0,0137	0,0160	0,0170	0,0184	0,0335	0,0329

Tabela 3.6 - Tempos de execução em segundos das Medidas de Validação para o conjunto das amostras X2 particionado pelo FCM para 2 a 10 categorias nebulosas e m = 2. Os índices 1 a 4 indicam a posição das medidas de validação em relação ao tempo de execução.

Outra característica importante da medida *ICC* é que ela também pode ser empregada para *validar conjuntos rígidos*. Neste caso, a matriz *U* é uma matriz rígida de 1's e 0's, indicando que cada amostra pertence a um único conjunto. A amostra tem valor 1 se pertence à categoria e 0 se não pertence.

O método de categorização rígido *k-Means* (seção 2.1.1) foi aplicado para os conjuntos de amostras *X1* e *X2*, variando o número de categorias de 2 a 8.

Medidas		**Número de Categorias**						
		2	**3**	**4**	**5**	**6**	**7**	**8**
ICC	**M**	34,1800	77,8350	81,8485	**105,4463**	15,0987	14,8891	13,4127
DLF	**M**	0,7269	2,6561	5,9021	67,262	72,354	77,413	**79,549**
CS	**m**	0,3318	0,1350	0,1195	**0,0121**	0,6593	0,7413	16,1588

Tabela 3.7 - Valores das Medidas de Validação para o conjunto das amostras X1 particionado pelo método k-Means para 2 a 8 categorias rígidas

A medida *ICC* é então executada para as partições rígidas em comparação com o *DLF*, que é o critério *J*, definido pela eq. 2-50 e com o *CS*, que pode, como a *ICC*, validar conjuntos rígidos.

Pôde-se observar ao executar o método *k-Means* e analisar seus resultados que este tem maior tendência a cair em mínimos locais do que o método *FCM*. Um mínimo local para o conjunto de amostras *X1* é o centro global dos pontos.

Como a inicialização da matriz de funções características *U* do método *k-Means* é aleatória (assim como a matriz *U* do *FCM*), os centros tendem a ser posicionados em coordenadas próximas ao centro global do sistema. Deste modo, algumas categorias geradas não contem pontos e portanto os seus centros não são atualizados pelo algoritmo.

Tempos	**Número de Categorias**						
	2	**3**	**4**	**5**	**6**	**7**	**8**
ICC	0,0055	0,0066	0,0074	**0,0080**[1]	0,0085	0,0093	0,0102
DLF	0,6854	1,0458	1,3216	1,6784	2,0324	2,3002	2,6140
CS	0,0184	0,0244	0,0308	**0,0377**[2]	0,0437	0,0502	0,0569

Tabela 3.8 - Tempos em segundos das Medidas de Validação para o conjunto das amostras X1 particionado pelo método k-Means para 2 a 8 categorias rígidas. Os índices 1 e 2 indicam a posição das medidas de validação em relação ao tempo de execução.

Este problema não acontece com o método *FCM*, dado que todos os pontos pertencem a todas as categorias geradas.

Nas partições rígidas geradas, o comportamento crescente do *DLF* (Tabela 3.7) pode ser observado mesmo no conjunto de amostras *X1*, onde não há sobreposição das classes. Para este conjunto de dados, o *DLF* erra, validando o maior número de partições calculado, que é o 8, como o melhor número de partições, ao invés de 5 categorias.

Medidas		**Número de Categorias**						
		2	**3**	**4**	**5**	**6**	**7**	**8**
ICC	**M**	11,3917	11,9251	15,5823	**18,1940**	13,4461	13,3913	14,9289
DLF	**M**	1,1355	1,8411	2,9176	4,8258	5,4257	6,0781	**6,8428**
CS	**m**	0,2177	0,3326	0,2488	**0,1898**	0,3928	0,4338	0,3717

Tabela 3.9 - Valores das Medidas de Validação para o conjunto das amostras X2 particionado pelo método k-Means para 2 a 8 categorias rígidas

A *ICC* continua sendo a medida mais rápida, tanto para o conjunto de amostras *X1* como para o *X2*, seguida pelo *CS* e pelo *DLF*, nesta ordem, como pode ser observado nas tabelas de tempo de execução Tabela 3.8 e Tabela 3.10.

Tempos	**Número de Categorias**						
	2	**3**	**4**	**5**	**6**	**7**	**8**
ICC	0,0050	0,0066	0,0074	**0,0099**[1]	0,0102	0,0115	0,0135
DLF	0,7360	1,0062	1,3258	1,6534	1,9850	2,3288	2,6166
CS	0,0187	0,0247	0,0321	**0,0382**[2]	0,0454	0,0516	0,0582

Tabela 3.10 – Tempos em segundos das Medidas de Validação para o conjunto das amostras X2 particionado pelo método k-Means para 2 a 8 categorias rígidas. Os índices 1 e 2 indicam a posição das medidas de validação em relação ao tempo de execução.

Esses experimentos indicam que a medida proposta *ICC* é uma medida de validação rápida e eficiente quando aplicada na avaliação tanto de partições nebulosas como de partições rígidas.

Seu comportamento mostrou que ela é capaz de analisar com eficiência as partições e acertar o número ótimo destas, mesmo quando as outras medidas de validação falham devido à alta sobreposição das amostras.

Comparada às medidas de validação que não são sensíveis à sobreposição das amostras, ela provou ser uma medida mais rápida sendo assim mais apropriada para ser aplicada a grandes sistemas, diminuindo o custo computacional destes.

4 ICC-KNN : UM SISTEMA ESTATÍSTICO NÃO-PARAMÉTRICO DE RECONHECIMENTO DE PADRÕES

Os métodos de categorização nebulosos *FCM*, *GK* e *GG* (seções 2.2.1, 2.2.2 e 2.2.3) são métodos que conseguem analisar e distinguir aglomerados de amostras de *formas convexas* que sejam linearmente separáveis.

Quanto ao formato das classes, o método *FCM* consegue distinguir bem classes hiperesféricas, o método *GK* distingue classes hiperelipticas que tenham exatamente o mesmo número de elementos, enquanto que o método *GG* consegue distinguir classes com *formas convexas* indefinidas.

Mas quando as classes de um problema apresentam *formas côncavas* ou não são contínuas, estes métodos não conseguem mapear cada classe em uma única categoria, agrupando apenas parte dos pontos das classes na categoria ou agrupando os pontos da classe côncava junto com pontos de outras classes.

O método *K-NN nebuloso* (seção 2.2.5) consegue lidar melhor com os vários formatos que as classes de um problema podem assumir. Mas para que a classificação gerada por ele obtenha um bom resultado, é necessário escolher os *padrões* que melhor representam cada classe.

Neste capítulo será apresentada a proposta de um *Sistema Estatístico Não-Paramétrico de Reconhecimento de Padrões*, o *Sistema ICC-KNN*, que associa as vantagens dos métodos de categorização e classificação nebulosos *FCM* e *K-NN nebuloso*, a fim de lidar com dados dispostos em diversos formatos de classes.

4.1 SISTEMA ICC-KNN

O *Sistema ICC-KNN* (Figura 4.1) é um *Sistema Estatístico Não-Paramétrico de Reconhecimento de Padrões* que emprega em seu *Módulo de Modelagem* dois métodos de categorização nebulosos consagrados, o *FCM* e o *K-NN nebuloso* em conjunto com a medida de validação *ICC* proposta nesta dissertação e em seu *Módulo de Reconhecimento de Padrões*, o método *K-NN nebuloso*. (FRANCO, 2002)

O papel do *Módulo de Modelagem* no sistema é estabelecer as estruturas dos dados, enquanto que o papel do *Módulo de Reconhecimento de Padrões* é receber os dados não classificados e atribuí-los às classes definidas no processo de classificação.

O *Módulo de Modelagem* do sistema *ICC-KNN* avalia, a partir de dados previamente conhecidos, quais os *melhores padrões*, o *melhor número de vizinhos* e a melhor *constante nebulosa* a serem utilizados como parâmetros para o método *K-NN nebuloso*.

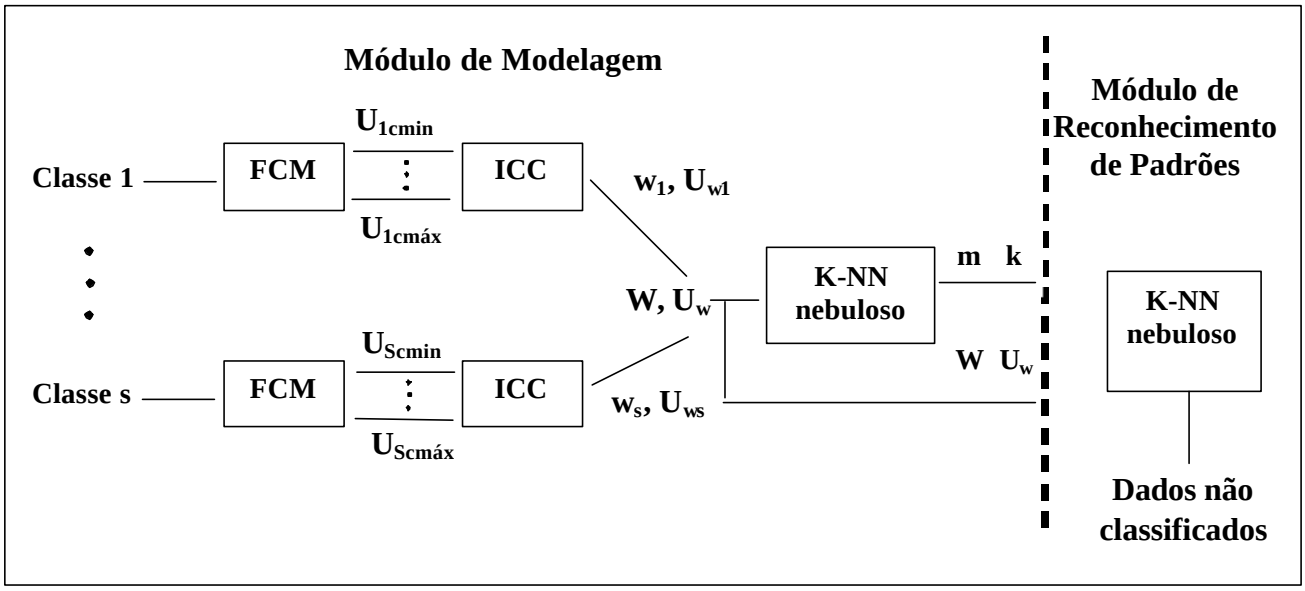

Figura 4.1 - Representação do Sistema ICC-KNN. W é o vetor de padrões escolhido a partir dos valores máximos da medida de validação ICC, gerados a partir da validação das execuções do método de categorização FCM. A saída do algoritmo é formada pela a matriz U e pelos valores de k, e m que obtiveram maior taxa de acerto rígido.

O conhecimento prévio do número de classes de um problema não implica que o número de categorias de um método de categorização seja conhecido, pois na maioria

das vezes, uma categoria não consegue representar exatamente uma classe, como no caso de *classes côncavas* ou *não-contínuas*.

Para estes tipos de conjuntos de amostras, o *K-NN nebuloso* é o melhor método para reconhecer classes, dado que ele é um método supervisionado que pode distinguir classes com diferentes formatos.

Os métodos mais usados para a escolha de padrões para o *K-NN nebuloso* são a *escolha aleatória* de pontos da massa de dados ou a *escolha por inspeção visual*.

Porém escolher aleatoriamente os padrões não significa necessariamente escolher os melhores padrões que representam cada classe, enquanto que escolher padrões visualmente é um método trabalhoso e não muito preciso, sem contar que as classes devem ter uma representação visual para este tipo de escolha, o que não acontece na maioria dos casos.

Outro método de escolha comum é a utilização de todo o conjunto de amostras conhecido, ou grande parte dele, como padrões das classes. Porém o custo computacional desta escolha é muito grande, dado que o tempo de execução do algoritmo fica extremamente alto.

O sistema *ICC-KNN* propõe escolher os padrões do método *K-NN nebuloso* através da aplicação do método *FCM* para cada classe do problema separadamente, executando-o para um intervalo de categorias fixo.

O método *FCM* foi escolhido por sua capacidade de posicionar os centros das categorias onde há maior concentração de amostras, representado assim cada diferente localização espacial dos aglomerados hiperesféricos dos pontos de cada classe.

Assim, os centros gerados pelo método *FCM* são propostos nesta dissertação como bons candidatos a padrões para o método *K-NN nebuloso*.

Após executar o *FCM*, é necessário saber qual o melhor número de centros que representa cada classe. O número ideal de categorias que representa cada classe é então encontrado através da validação feita pela medida *ICC* proposta na seção 3.4.

O *Módulo de Modelagem* do sistema *ICC-KNN* é dividido em duas fases de treinamento. A primeira fase consiste em encontrar a melhor distribuição de padrões para o *K-NN nebuloso*. A segunda fase consiste em afinar os parâmetros *m* e *k* do *K-NN nebuloso*.

Na *primeira fase* do treinamento, o método *FCM* é aplicado aos dados de treinamento de cada classe separadamente. Para isto, devem ser fixados a *constante nebulosa m* e o intervalo de categorias nebulosas [c_{min} , $c_{máx}$] para o qual o *FCM* será executado.

Neste intervalo, c_{min} e $c_{máx}$ são os valores mínimo e máximo escolhidos de *c*, para c_{min} maior ou igual a 2 e $c_{máx}$ menor que o número de pontos da classe que contém menos pontos, dado que c_{min} e $c_{máx}$ serão iguais para todas as classes. Os centros calculados em cada execução do *FCM* neste intervalo para cada classe são candidatos a padrões de suas respectivas classes no método *K-NN nebuloso*.

A seguir, a medida de validação *ICC* é utilizada para validar a melhor execução do método *FCM* para cada classe. Os centros da matriz de centros *V*, associada à matriz de graus de inclusão *U* que maximiza o valor da medida *ICC*, são escolhidos como os padrões de sua classe. Assim são definidos quantos e quais padrões representarão cada classe no *K-NN nebuloso*.

O relacionamento entre os padrões e suas classes é definido pelos seus graus de inclusão em cada classe. Para o sistema *ICC-KNN*, as partições dos padrões são rígidas, ou seja, cada padrão representa uma única categoria na matriz de graus de inclusão dos padrões U_W, tendo *grau de inclusão* 1 para a categoria que representa e 0 para as demais.

Os padrões são rígidos porque o conjunto original de dados é rígido, ou seja, cada ponto pertence somente à sua classe, o que permite que cada classe seja analisada separadamente e os padrões sejam calculados separadamente.

Na *segunda fase* do treinamento, a constante nebulosa *m* e o número de vizinhos *k* do método *K-NN nebuloso* são variados a fim de encontrar os valores que obtém a melhor performance do método.

Os valores de *m* e de *k* são variados nos intervalos [m_{min} , $m_{máx}$] e [k_{min} , $k_{máx}$] respectivamente, onde m_{min} e $m_{máx}$ são os valores mínimo e máximo de *m* e k_{min} e $k_{máx}$ são os valores mínimo e máximo de *k* escolhidos.

O *K-NN nebuloso* é então executado para cada valor *m* e *k* destes intervalos. Cada ponto amostral apresentado ao algoritmo é então associado à classe em que tem maior grau de inclusão. Se o ponto for associado à sua própria classe, houve um acerto de classificação, caso contrário ocorreu um erro.

A taxa de *acertos rígidos* calcula a percentagem de amostras que foram classificadas corretamente como pertencentes a sua classe de origem. Os valores de *m* e *k* que obtêm maior taxa de *acertos rígidos* são escolhidos como os melhores parâmetros para o *K-NN nebuloso*.

Caso dois ou mais valores diferentes de *k* obtenham a mesma taxa de *acertos rígidos*, o menor valor de *k* é escolhido, independente do valor de *m*. É vantajoso escolher o menor *k* em caso de empate, dado que o tempo de execução cresce conforme o número de vizinhos aumenta.

Se dois ou mais valores iguais de *k* obtiverem a mesma taxa de *acertos rígidos*, o número de vizinhos associado à menor *constante nebulosa* é escolhido, dado que o valor de *m* não interfere no custo computacional e que as categorias ficam mais rígidas, diminuindo a sobreposição entre estas.

No *Módulo de Reconhecimento de Padrões*, o método *K-NN nebuloso* é aplicado aos dados a serem classificados, utilizando os padrões e os parâmetros avaliados no *Módulo de Modelagem.*

Para o **Algoritmo ICC-KNN** apresentado abaixo, seja $R = \{r_1, \ldots, r_n\}$ um conjunto de *n* amostras rotuladas do espaço $\Re^p$ pertencentes a *s* classes conhecidas, U_{sc} e V_{sc} a matriz de graus de inclusão e a matriz de centros, respectivamente, geradas pelo método *FCM* para a classe *s* com *c* categorias e $W = \{w_1, \ldots, w_s\} \in \Re^p$ o conjunto dos conjuntos dos padrões w_i das *s* classes.

Algoritmo ICC-KNN:

- **MÓDULO DE MODELAGEM**

Primeira fase do Treinamento

Passo 1. Fixar *m*
Passo 2. Fixar c_{min} e $c_{máx}$

Passo 3. Para cada classe *s* conhecida
Gerar o conjunto R_s com os pontos de *R* pertencentes à classe s
Para cada categoria c no intervalo $[c_{min}, c_{máx}]$
Executar *FCM* para c e o conjunto R_s gerando U_{sc} e V_{sc}
Calcular a *ICC* para R_s e U_{sc}
Fim

Definir os padrões w_s da classe s como a matriz V_{sc} que maximiza a *ICC*
Fim

Passo 4. Gerar o conjunto W = {w_1, ..., w_s}

Segunda fase do Treinamento

Passo 5. Fixar m_{min} e $m_{máx}$

Passo 6. Fixar k_{min} e $k_{máx}$
Para cada *m* do intervalo [m_{min} , $m_{máx}$]
Para cada *k* do intervalo [k_{min} , $k_{máx}$]
Executar o *K-NN nebuloso* para os padrões do conjunto *W*, gerando U_{mk}
Calcular os acertos rígidos para U_{mk}
Fim
Fim

Passo 7. Escolher o *m* e *k* que obtêm a maior taxa de *acertos rígidos*

Passo 8. Se houver empate
Se os *k* são diferentes
Escolher o menor *k*
Senão
Escolher o menor *m*
Fim
Fim

- **MÓDULO DE RECONHECIMENTO DE PADRÕES**

Passo 9. Aplicar o *K-NN nebuloso* com os padrões do conjunto W e os parâmetros *m* e *k* escolhidos aos dados a serem classificados

4.2 AVALIANDO O SISTEMA ICC-KNN

Para avaliar e testar o sistema *ICC-KNN* foi criado um conjunto de 2000 amostras bidimensionais dispostas em quatro classes de 500 pontos cada, como pode ser visto na Figura 4.2.

Observa-se que a *classe 1* e a *classe 4* têm formato côncavo, onde a *classe 1* tem o formato da letra *S* e a *classe 4* o formato da letra *C*. A *classe 2* e a *classe 3* são classes convexas de formato elíptico.

As classes 2 e 3 não ofereceriam dificuldade de categorização pelos métodos *FCM*, *FKCN*, *GG* ou *GK* se apresentadas sem as outras duas classes. Entretanto a

presença das classes côncavas 1 e 4 prejudica a habilidade destes métodos em reconhecê-las.

As 2000 amostras foram divididas em *dados de treinamento* e *dados de teste*. Os *dados de treinamento* perfazem 80% das amostras, contendo 400 pontos de cada classe e os *dados de teste* perfazem 20% das amostras, com 100 pontos de cada classe.

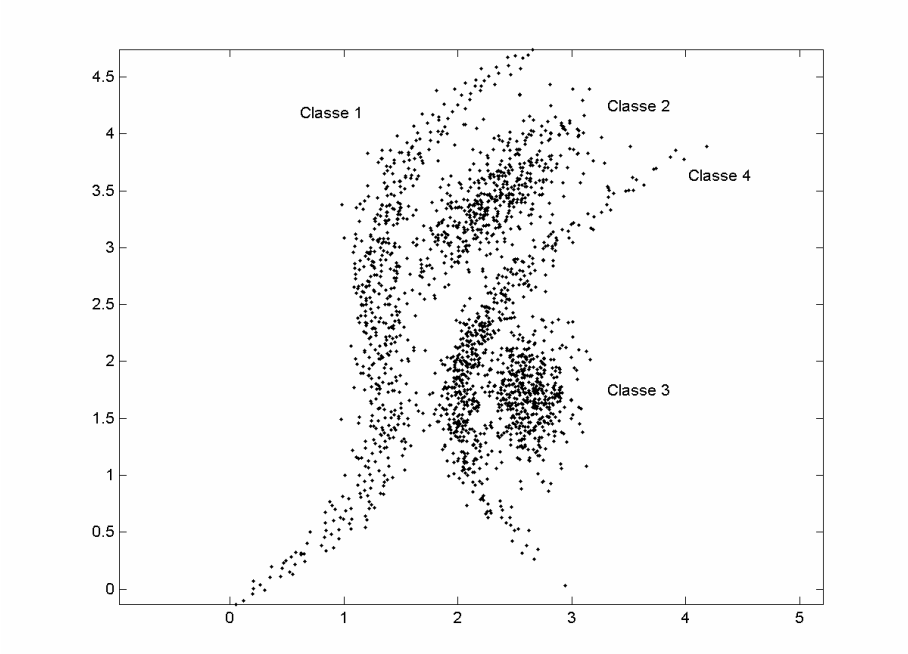

Figura 4.2 – Conjunto de 2000 amostras bidimensionais dispostas em quatro classes. As classes 1 e 4 têm um formato côncavo, a classe 2 tem um formato elíptico e a classe 3 tem um formato aproximadamente circular.

Na *primeira fase do treinamento – PFT* do *Módulo de Modelagem* do sistema *ICC-KNN,* os *dados de treinamento* foram divididos em suas respectivas classes. Cada classe foi apresentada separadamente ao método *FCM*, que foi executado com o número de categorias *c* variando de 3 a 7 para a constante nebulosa *m* fixa e igual a *1,25*.

A medida de validação *ICC* identificou as melhores partições geradas pelo *FCM* como *4 categorias* para as classes 1 e 4 e *3 categorias* para as classes 2 e 3. Os centros destas categorias serão utilizados como padrões de suas respectivas classes no método *K-NN nebuloso*.

Na *segunda fase de treinamento – SFT* do *Módulo de Modelagem,* o método *K-NN nebuloso* foi executado para os dados de treinamento, com os padrões (centros)

gerados na *PFT*, variando o valor de *k* de 3 à 7 vizinhos para os valores de $m \in \{1,1;\ 1,25\ ;\ 1,5;\ 2\}$.

Os valores da constante nebulosa $m = 1,25$ e $m = 2$ são os mais comumente utilizados na literatura. O valor de $m = 1,1$ foi utilizado para analisar o comportamento de partições mais próximas de partições rígidas, dado que quanto menor o valor de *m*, mais rígidas são as partições até chegar ao valor $m = 1$, onde as partições são totalmente rígidas.

O valor de $m = 1,5$ foi utilizado para analisar o comportamento das partições entre os valores mais utilizados de *m*.

Para mostrar a melhoria na capacidade de classificação obtida pelo sistema *ICC-KNN* com utilização dos padrões escolhidos na *PFT*, a *SFT* foi repetida utilizando padrões escolhidos aleatoriamente.

Os *padrões aleatórios* são amostras extraídas aleatoriamente dos dados de treinamento para cada classe. O número de padrões aleatórios para cada classe foi o mesmo validado pela medida de validação *ICC*.

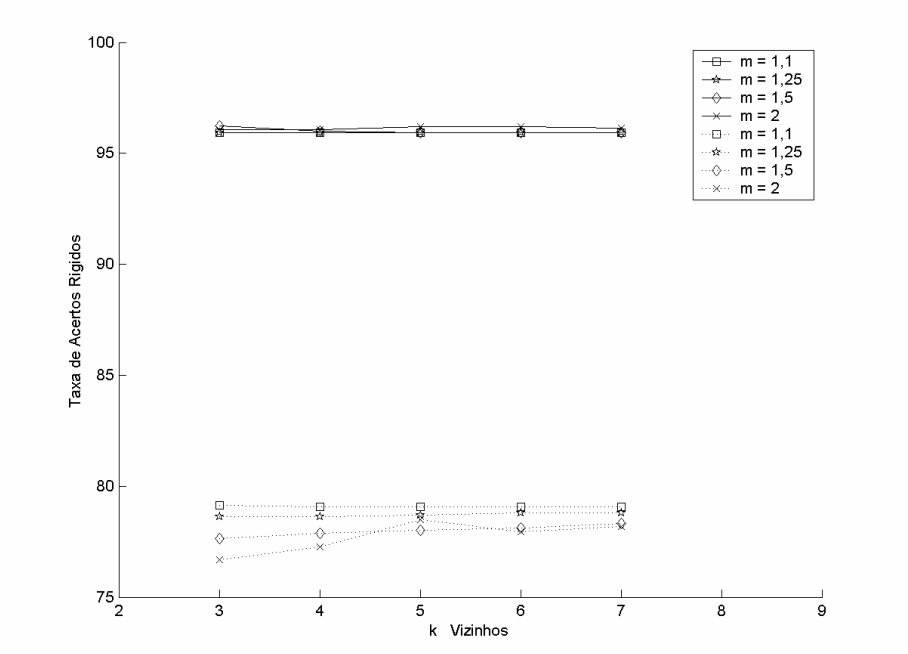

Figura 4.3 – Gráfico da Taxa de Acertos Rígidos em função do número de vizinhos k e da constante nebulosa m. A linha sólida representa os acertos do método K-NN nebuloso para os padrões da PFT e a linha pontilhada representa os acertos para os padrões escolhidos aleatoriamente

O gráfico na Figura 4.3 mostra os resultados da comparação descrita acima. Ele apresenta a taxa de *acertos rígidos* em função do número de *vizinhos k* e da *constante nebulosa m*.

A *linha sólida* mostra as taxas de *acertos rígidos* obtidas internamente na *SFT* do sistema *ICC-KNN* com os padrões gerados na *PFT* e a *linha tracejada* apresenta as taxas de *acerto rígido* obtidas para os padrões escolhidos aleatoriamente.

Pode-se observar que as taxas de *acerto rígidos* para os padrões gerados na *PFT* são muito semelhantes entre si para este problema, sendo praticamente independentes do valor da *constante nebulosa m* e das variações de *k*.

Já as taxas de *acertos rígidos* para os padrões aleatórios mostram maior dependência dos valores de *m*, indicando que o *K-NN nebuloso* é mais estável em relação ao valor de *m* para os padrões obtidos na *PFT* do que para os padrões escolhidos aleatoriamente.

As *matrizes de confusão* dos melhores resultados para os padrões escolhidos na *PFT* e para os padrões aleatórios são ilustradas na Tabela 4.1.

O sistema *ICC-KNN*, em seu *Módulo de Modelagem*, encontrou para $m = 1,5$ e $k = 3$ a maior taxa de *acertos rígidos* do método *K-NN nebuloso* com os padrões obtidos na *PFT*, que foi de 96,25%. Para os padrões aleatórios, a *SFT* do sistema obteve maior taxa de *acertos rígidos* para $m = 1,1$ e $k = 3$, que foi de 79,13%.

Este resultado demonstra a validade da utilização do sistema *ICC-KNN* para obter os padrões do método *K-NN nebuloso*, dado que a taxa de acertos obtida com os padrões da *PFT* foi 21,64% superior à obtida com os padrões aleatórios.

Dados de Treinamento								
Classes	**Padrões da PFT**				**Padrões Aleatórios**			
	1	**2**	**3**	**4**	**1**	**2**	**3**	**4**
1	388	10	0	2	213	66	0	121
2	14	379	0	7	19	380	0	1
3	0	0	376	24	3	0	324	73
4	0	1	2	397	4	46	1	349

Tabela 4.1 – Matriz de confusão para os melhores resultados do método K-NN nebuloso, para os padrões da PFT, com taxa de acertos de 96,25% para m = 1,5 e k = 3 e para os padrões aleatórios com taxa de acertos de 79,13% para m = 1,1 e k = 3. As linhas correspondem às classes e as colunas correspondem as amostras classificadas pelos métodos. As células sombreadas correspondem ao número de amostras que foram classificadas corretamente em suas respectivas classes.

Após a execução da *SFT* do *Módulo de Modelagem* do sistema, foram encontrados os melhores parâmetros a serem utilizados na execução do método *K-NN nebuloso* para os dados de teste, que são $m = 1,5$ e $k = 3$ no *Módulo de Reconhecimento de Padrões*.

O método *K-NN nebuloso* é então aplicado com estes parâmetros aos dados de teste, para os padrões computados na *PFT* e para os padrões escolhidos aleatoriamente. Os dados de teste são compostos de 100 pontos de cada classe.

Para os padrões calculados na *PFT*, a taxa de *acertos rígidos*, que foi de 94,75%, é bem maior que a taxa obtida pelo método com os padrões aleatórios, que foi de 79%.

Este resultado confirma a validade do sistema *ICC-KNN* para calcular os padrões a serem utilizados pelo *K-NN nebuloso*, dado que para os mesmos parâmetros *m* e *k* calculados na *SFT*, o método aplicado aos padrões aleatórios obteve uma taxa inferior do que quando aplicado aos padrões calculados na *PFT*.

m = 1,5 e k = 3	**Classe 1**	**Classe 2**	**Classe 3**	**Classe 4**
Pad. da PFT	96,08%	98,94%	93,88%	88,68%
Pad. Aleat.	95,40%	80,17%	83,78%	84,88%

Tabela 4.2 - Percentagem de pontos que pertencem a cada classe e foram classificados como tais, em relação ao total de pontos classificados como pertencentes a cada classe, geradas pelo método K-NN nebuloso aplicado aos dados de teste para os padrões gerados na primeira fase de treinamento e para os escolhidos aleatoriamente, com m = 1,5 e k = 3.

O ganho em utilizar o *K-NN nebuloso* com os padrões da *PFT* em relação aos padrões aleatórios na fase de teste foi de 20%.

Dados de Testes								
Classes	**Padrões da PFT**				**Padrões Aleatórios**			
	1	**2**	**3**	**4**	**1**	**2**	**3**	**4**
1	97	2	0	1	53	27	0	20
2	4	93	0	3	4	96	0	0
3	0	0	90	10	0	0	82	18
4	0	0	1	99	0	15	0	85

Tabela 4.3 – Matriz de confusão para a execução do K-NN nebuloso para os dados de teste, com m = 1,5 e k = 3, para os padrões da PFT e para os escolhidos aleatoriamente. As linhas correspondem às classes e as colunas correspondem às amostras classificadas pelo método. As células sombreadas correspondem ao número de amostras que foram classificadas corretamente em suas respectivas classes

Tabela 4.2 mostra dentre todos os pontos que foram classificados como pertencentes a cada classe, a percentagem de pontos que realmente pertencem a esta classe e foram classificados como tais e a Tabela 4.3 ilustra as *matrizes de confusão* para os resultados obtidos com os dados de teste, para $m = 1{,}5$ e $k = 3$.

Para os padrões calculados na *PFT*, foi a classe 2 que agrupou mais pontos de sua própria classe, enquanto que, para os padrões aleatórios, foi a classe 1.

4.2.1 TEMPOS DE EXECUÇÃO

O tempo de computação para a *PFT* do sistema *ICC-KNN*, onde são calculados os padrões de cada classe através da execução do método *FCM* e da medida de validação *ICC*, foi de 15,5 segundos.

A *SFT*, onde são variados os valores de *m* e de *k* na execução do método *K-NN nebuloso* e são calculadas as taxas de *acertos rígidos*, durou 21,04 segundos. O *tempo total de treinamento* do sistema *ICC-KNN* foi de aproximadamente 36,5 segundos.

O *K-NN nebuloso* executado com os padrões aleatórios não efetua os passos da *PFT*. Ao invés disto, ele recebe a informação do número de padrões de cada classe, que é igual à já calculada na *PFT*, e extrai os padrões aleatoriamente dos dados de treinamento para cada classe.

Após isto, ele executa os passos da *SFT*. Seu tempo de treinamento é então menor que o do sistema *ICC-KNN*, sendo de 23,11 segundos, porém suas taxas de acerto também são menores. Os tempos de treinamento e as taxas de acertos dos métodos são mostrados na Tabela 4.4 da seção 4.4.

4.3 TAXA DE ACERTOS NEBULOSOS

A qualidade da classificação obtida pelo sistema *ICC-KNN* também pode ser medida em função da proximidade dos pontos amostrais aos padrões de sua classe de origem, ou seja, através da análise dos *graus de inclusão* dos pontos nas classes do problema.

A medida criada nesta dissertação para avaliar esta qualidade é denominada *acertos nebulosos*. Esta medida calcula os acertos nebulosos do método a partir da

matriz de graus de inclusão *U*. Ela não faz parte do sistema *ICC-KNN*, mas foi incluída nesta dissertação para mostrar uma análise nebulosa do problema.

Uma amostra é classificada corretamente no sentido nebuloso se o *grau de inclusão* na classe a qual ela pertence é maior que o *inverso do número de vizinhos 1/k* utilizados no *K-NN nebuloso*.

Se o *grau de inclusão* for igual em todas as classes, significa que não houve realmente uma classificação, dado que a amostra foi classificada como pertencente igualmente a todas as classes.

Se o *grau de inclusão* for menor que *1/k* na sua classe de origem, significa que a amostra teve maior afinidade com uma classe que não é a sua classe original. Estes dois casos são considerados erros de classificação nebulosos.

A taxa de *acertos nebulosos* pode ser utilizada no sistema *ICC-KNN* em substituição à taxa de *acertos rígidos*, tendo as duas a mesma função na avaliação dos melhores parâmetros para o método *K-NN nebuloso*.

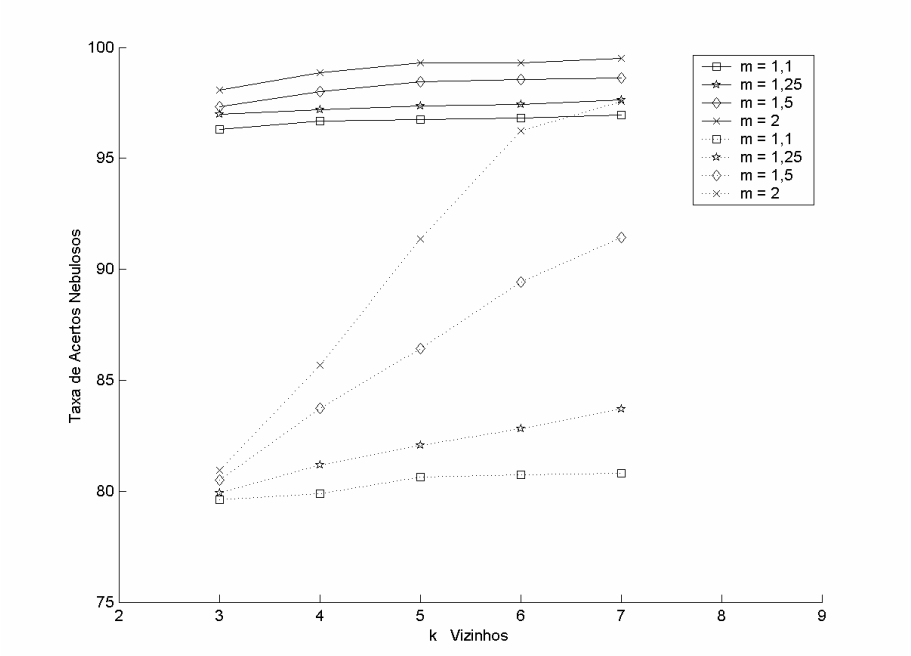

Figura 4.4 – Gráfico da Taxa de Acertos Nebulosos em função do número de vizinhos k e da constante nebulosa m. A linha sólida representa os acertos do método K-NN nebuloso para os padrões calculados na primeira fase de treinamento e a linha pontilhada representa os acertos para os padrões escolhidos aleatoriamente

A única diferença entre elas é que ao calcular os *acertos nebulosos*, todos os *graus de inclusão* obtidos pelo *K-NN nebuloso* são considerados, obtendo-se assim, uma visão do estado nebuloso do sistema.

Na Figura 4.4, a *linha sólida* mostra as taxas de *acertos nebulosos* obtidas internamente na *SFT* do sistema *ICC-KNN* com os padrões gerados pela *PFT* e a *linha tracejada* apresenta as taxas de *acertos nebulosos* obtidas para os padrões escolhidos aleatoriamente para as classes da Figura 4.2.

As taxas de *acertos nebulosos* também são mais estáveis em relação à constante nebulosa para os padrões calculados na *PFT* do que para os padrões escolhidos aleatoriamente.

Para partições mais nebulosas, com $m = 1{,}5$ e $m = 2$, as taxas de *acertos nebulosos* para os padrões aleatórios sofrem grandes variações em função do aumento do número de vizinhos. Para as partições mais rígidas, esta variação é bem pequena.

A *taxa de acertos nebulosos*, como a de *acertos rígidos*, é maior para o método *K-NN nebuloso* executado com os padrões calculados na *PFT*, com 95,75% de acertos, do que para a execução com padrões aleatórios, com 83% de acertos. Estas taxas podem ser comparadas na Tabela 4.4 da seção 4.4.

4.4 COMPARANDO O SISTEMA ICC-KNN COM OS MÉTODOS DE CATEGORIZAÇÃO ESTUDADOS

Para mostrar a eficiência do sistema *ICC-KNN* sobre outros métodos de categorização, os métodos *FCM*, *FKCN*, *GG* e *GK* foram executados para os mesmos conjuntos de treinamento e de teste descritos acima, para $c = 4$ categorias, dado que o problema tem 4 classes reais, e para os valores da constante nebulosa $m \in \{1{,}1;\ 1{,}25;\ 1{,}5;\ 2\}$.

Como estes métodos apenas categorizam os dados sem classificá-los, foi necessário criar nesta tese alguns passos complementares a fim de obter uma classificação final.

Estes passos, juntamente com os métodos de categorização, também formam um *Sistema Estatístico Não-Paramétrico de Reconhecimento de Padrões* para cada método de categorização utilizado (seção 1.3), com *fases de treinamento* e de *teste*.

Na *fase de treinamento*, os *dados de treinamento* foram fornecidos aos métodos. Os métodos *FCM*, *GG* e *GK* foram executados nesta fase, gerando os centros das 4 categorias que minimizavam a função objetivo (eq. 2-8) para suas respectivas distâncias.

O método *GG* foi executado com os centros gerados pelo método *FCM* e com inicialização aleatória dos *graus de inclusão*. A rede neural *FKCN* foi treinada até encontrar sua melhor configuração.

O próximo passo foi associar as categorias geradas às classes reais do problema. Os métodos de categorização não fornecem um meio de associar as categorias geradas às classes do problema, sendo necessário criar um critério que faça esta associação.

O *critério do somatório dos graus de inclusão* foi então proposto nesta dissertação para gerar esta associação. Este critério é baseado no cálculo dos somatórios dos *graus de inclusão* dos pontos de cada classe do problema em cada categoria gerada.

Como os métodos são nebulosos, para cada categoria é calculado o somatório dos *graus de inclusão* dos pontos de treinamento de cada classe do problema nesta categoria. Assim, uma categoria representa a classe cujos pontos obtiveram maior grau de inclusão nesta.

Deste modo, uma classe pode ser representada por uma ou mais categorias, implicando no desaparecimento de classes no resultado final. Porém isto é um efeito colateral aceitável, pois permite a obtenção de maiores taxas de acertos.

Na segunda fase, chamada de *fase de teste*, os métodos foram novamente executados, agora para os dados de testes. Os métodos *FCM*, *GG* e *GK* foram inicializados com os centros gerados na *fase de treinamento* e só executaram um passo para calcular o *grau de inclusão* dos pontos de teste em cada categoria, sem remanejar os centros.

Após a execução, deve-se calcular o *grau de inclusão* dos pontos nas classes do problema. Se cada categoria corresponde a somente uma classe, o *grau de inclusão* do ponto na classe é igual ao *grau de inclusão* do ponto na categoria correspondente.

Foi necessário então criar também um critério para calcular o *grau de inclusão* de um ponto em uma classe representada por mais de uma categoria. Este critério implica em gerar este *grau de inclusão* a partir da soma dos *graus de inclusão* dos pontos nas categorias que representam a classe.

Métodos	ICC-KNN	K-NN Neb. Alea.	FCM	FKCN	GG	GK
Acertos Ríg.	94,75 %	79 %	66 %	66 %	69 %	84 %
Acertos Neb.	95,75 %	83 %	70,75 %	70,75 %	69 %	89,5 %
Tempos	36,5 s	23,11 s	2,91 s	2,59 s	22,66 s	18,14 s

Tabela 4.4 – Acertos rígidos, acertos nebulosos e tempo de treinamento dos métodos de categorização ICC-KNN, K-NN nebuloso com padrões aleatórios, FCM, FKCN, GG e GK

A Figura 4.5 mostra a taxa de *acertos rígidos* obtida pelos métodos ao categorizar os dados de teste. As taxas de *acertos rígidos* também são calculadas a partir da associação dos pontos às classes em que têm máximo grau de inclusão.

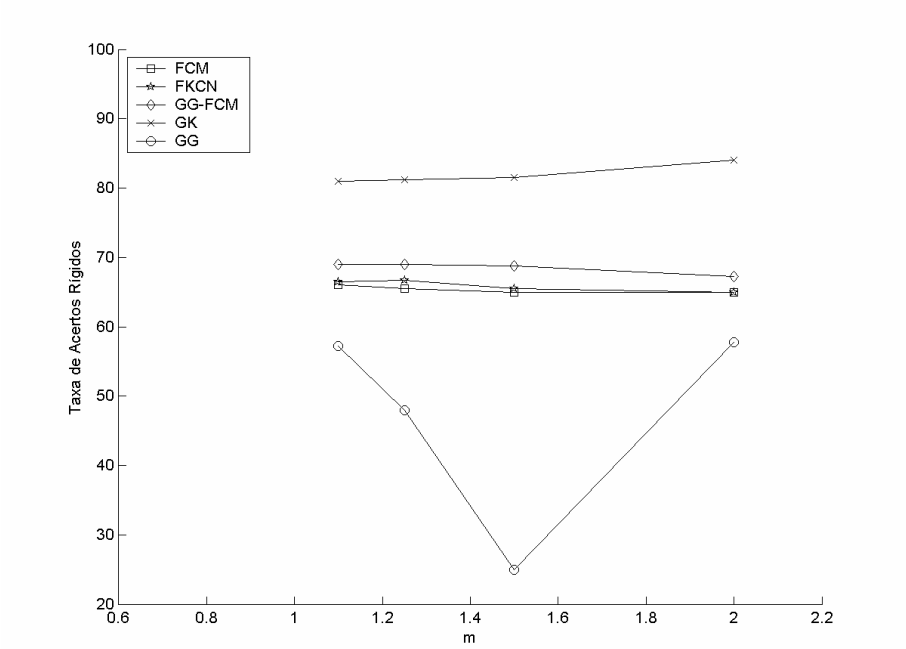

Figura 4.5 – Gráfico da Taxa de Acertos Rígidos dos métodos FCM, FKCN, GG e GK em função da constante nebulosa m para os dados de teste. GG-FCM é o método GG inicializado com os centros gerados pelo FCM e o GG é o método inicializado aleatoriamente.

A Tabela 4.4 resume as taxas de acertos rígidos e os tempos obtidos. A melhor taxa de *acertos rígidos* foi obtida pelo método *GK*, com 84% de acertos para $m = 2$.

Observa-se que esta taxa de acertos é inferior à melhor taxa de *acertos rígidos* obtida pelo sistema proposto *ICC-KNN*, que foi de 94,75%, o que confirma a vantagem do sistema *ICC-KNN* também sobre estes métodos de categorização quando utilizados para reconhecimento de padrões.

Em relação ao tempo de treinamento, o sistema *ICC-KNN* proposto é mais lento que os métodos *FCM*, *FKCN*, *GG* e *GK*, mas é da mesma ordem de grandeza destes. Os tempos e as taxas de acertos podem ser comparados através da Tabela 4.4.

O treinamento do sistema *ICC-KNN* para o dados analisados, foi apenas 18,36s mais lento que o método ganhador *GK*, uma diferença pequena que é compensada pelo seu melhor desempenho.

O método *GG*, quando inicializado aleatoriamente, pode obter taxas de acertos rígidos bem pequenas devido à sua tendência a má alocação dos centros. Seus resultados são inferiores aos obtidos quando há inicialização através dos centros gerados pelo *FCM*.

Sua melhor taxa de *acertos rígidos* foi de 57,75 % para m = 1,1 enquanto a pior taxa foi de 25% para m = 1,5. A pior taxa foi obtida quando os quatro centros gerados pelo método *GG* convergiram para um ponto de mesma coordenada.

Neste caso, somente uma categoria foi gerada pelo método. Esta categoria corresponde à classe que contém mais pontos. Como as classes do problema analisado têm o mesmo número de elementos, a categoria gerada corresponde à *classe 1*.

Assim todos os pontos da *classe 1* são classificados corretamente, enquanto os pontos das outras classes também são classificados como pertencentes à classe 1, gerando uma taxa de *acertos rígidos* de 25%.

A maior taxa para o método *GG* inicializado com os centros gerados pelo *FCM*, que aqui será chamado de *GG-FCM*, foi de 69% para m = 1,1 e m = 1,25, maior que a taxa obtida pelos métodos *FCM* e *FKCN*, porém menor que as taxas de acertos obtidas pelo método *GK*.

A taxa de *acertos rígidos* do método *FCM* e da rede neural *FKCM* são quase idênticas, o que é esperado, dado que os dois métodos utilizam a *distância Euclidiana* e que eles convergiram para centros muito próximos. Ambas tiveram suas maiores taxas de acertos empatadas em 66% para m = 1,1 e m = 1,25.

4.4.1 TAXA DE ACERTOS NEBULOSOS DOS MÉTODOS DE CATEGORIZAÇÃO

A qualidade da categorização dos métodos pode ser medida a partir da proximidade dos pontos aos centros de suas classes originais. Esta qualidade também pode ser analisada a partir dos *graus de inclusão* do ponto nas classes do problema.

A medida de qualidade criada nesta dissertação com este propósito também foi chamada de taxa de *acertos nebulosos* (seção 4.3) e é mostrada na Figura 4.6.

A única diferença para a mesma taxa definida para o sistema *ICC-KNN* é que ela é calculada a partir do *inverso do número de categorias 1/c* ao invés do inverso do número de vizinhos, medida inexistente para os métodos aqui testados.

A maior taxa de *acertos nebulosos* foi de 89,5%, tendo sido obtida pelo método *GK* para $m = 2$, indicando um bom grau de pertinência às suas classes originais.

Porém esta taxa foi bem menor que a obtida pelo método proposto *ICC-KNN*, que foi de 95,75%. Assim pode-se concluir que o sistema *ICC-KNN* obteve um grau de pertinência maior entre os pontos e suas respectivas classes.

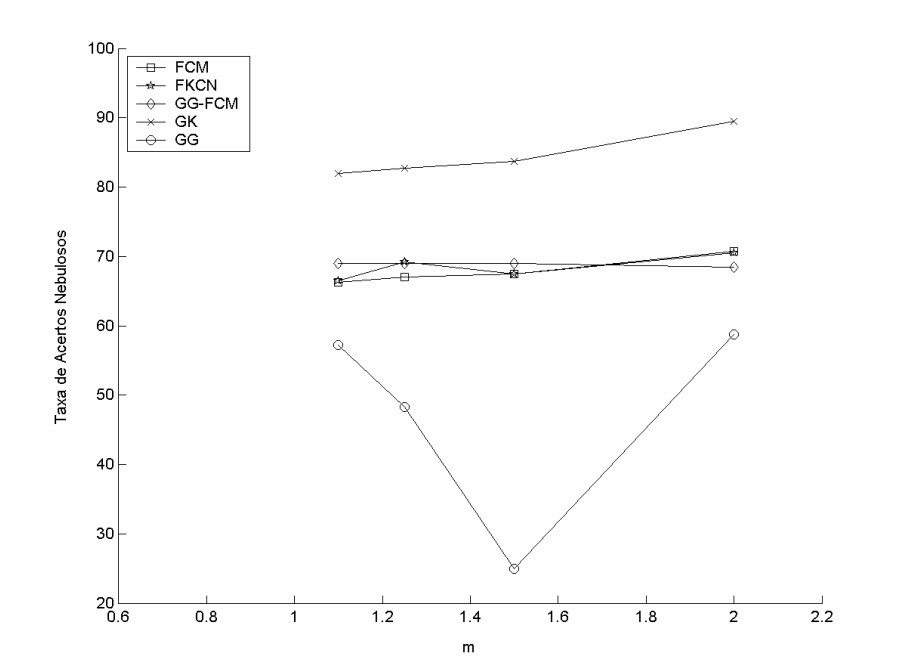

Figura 4.6 – Gráfico da Taxa de Acertos Nebulosos dos métodos FCM, FKCN, GG e GK em função da constante nebulosa m para os dados de teste. GG-FCM é o método GG inicializado com os centros gerados pelo FCM e o GG é o método inicializado aleatoriamente.

As taxas de *acertos nebulosos* são dependentes do valor da *constante nebulosa m*, com exceção do *GG-FCM*, cujas taxas ficaram em torno de 69% para todos os valores de *m*.

No geral, as taxas aumentaram conforme o valor de *m* aumentou. As segundas maiores taxa de *acertos nebulosos* foram obtidas pelos métodos *FCM* e *FKCN*, sendo de 70,75% para *m* = 2 nos dois métodos.

4.4.2 RESULTADOS GERADOS PELOS MÉTODOS DE CATEGORIZAÇÃO

A Figura 4.7 mostra a categorização gerada pelo método *FCM* nos dados de treinamento com *m* = 1,1, que obteve maior taxa de *acertos rígidos* para os dados de teste. É interessante observar o formato circular das categorias geradas.

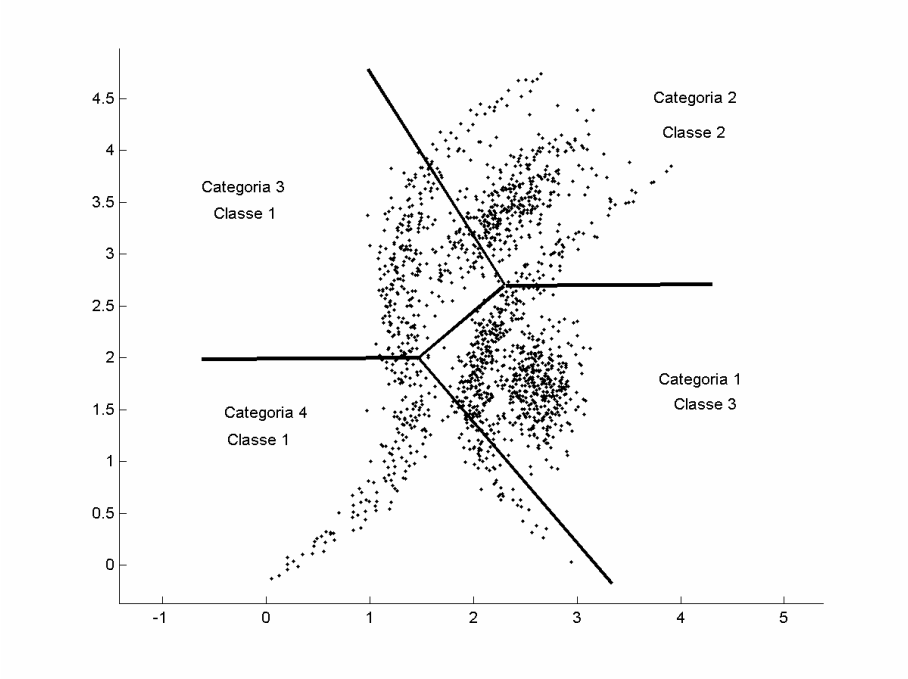

Figura 4.7 – Categorias geradas pelo algoritmo FCM para m = 1.1 na fase de treinamento para os dados de treino, e suas correspondentes classes

Nota-se que as 4 categorias geradas não correspondem exatamente às 4 classes reais do problema, pois cada categoria agrupou pontos de 2 ou mais classes.

Foi necessário então usar o critério do somatório dos *graus de inclusão* para relacionar cada categoria a uma classe a fim de classificar os dados de teste. Para facilitar o entendimento, as categorias serão apresentadas a partir dos pontos que agruparam.

A *categoria 1* agrupou pontos das classes 3 e 4, sendo a maioria dos pontos da classe 3. Logo a categoria 1 corresponde à *classe 3.* Isto significa que, na *fase de teste*, os pontos que foram agrupados na categoria 1 foram classificados como pertencentes à *classe 3*.

A *categoria 2* agrupou pontos das classes 1, 2 e 4, sendo a maioria dos pontos da classe 2, correspondendo, assim, à *classe 2*.

A *categoria 3* agrupou pontos das classes 1 e 2 e corresponde à classe 1. A *categoria 4* também corresponde à classe 1, pois agrupou pontos das classes 1 e 4, sendo a maioria de seus pontos pertencente à *classe 1*.

A *classe 1* é representada então por duas categorias, enquanto que a *classe 4* não é representada por nenhuma das categorias. Isto significa que nenhum ponto será classificado como pertencente à *classe 4*.

A Tabela 4.5 mostra a *matriz de confusão* da execução do método *FCM* para os pontos de teste. Pode-se observar na *matriz de confusão* que os pontos amostrais da *classe 4* foram distribuídos entre as outras 3 classes, obtendo taxa de acertos de 0%.

Classes	**FCM**			
	1	**2**	**3**	**4**
1	84	16	0	0
2	20	80	0	0
3	0	0	100	0
4	22	19	59	0

Tabela 4.5 – Matriz de confusão para a execução do FCM para os dados de teste para m = 1,1. As linhas correspondem às classes (valor esperado) e as colunas correspondem as amostras classificadas pelo método (valor estimado). As células sombreadas correspondem ao número de amostras que foram classificadas corretamente em suas respectivas classes

Também se observa que todos os pontos da *classe 3* foram classificados corretamente como pertencentes a esta classe.

A Tabela 4.6 representa dentre todos os pontos que foram classificados como pertencentes a cada classe, a percentagem de pontos que realmente pertencem a esta classe e foram classificados como tais.

Esta medida informa a qualidade da categorização gerada, dado que quanto mais pontos uma categoria tiver da classe que representa, melhor representada está a classe.

Classes	1	2	3	4
FCM	66,67%	69,57%	62,89%	0%
FKCN	75%	69,30%	57,47%	0%
GG-FCM	94,05%	71,85%	55,25%	0%
GK	92,77%	94%	75,19%	77,27%

Tabela 4.6 – Percentagem de pontos que pertencem a cada classe e foram classificados como tais, em relação ao total de pontos classificados como pertencentes a cada classe, para os métodos FCM, FKCN, GG inicializado pelos centros gerados pelo método FCM (GG-FCM), GG e GK.

A categorização gerada pela rede neural *FKCN* para os dados de treinamento com m = 1,1, que obteve maior taxa de acertos rígidos para os dados de teste, é ilustrada na Figura 4.8.

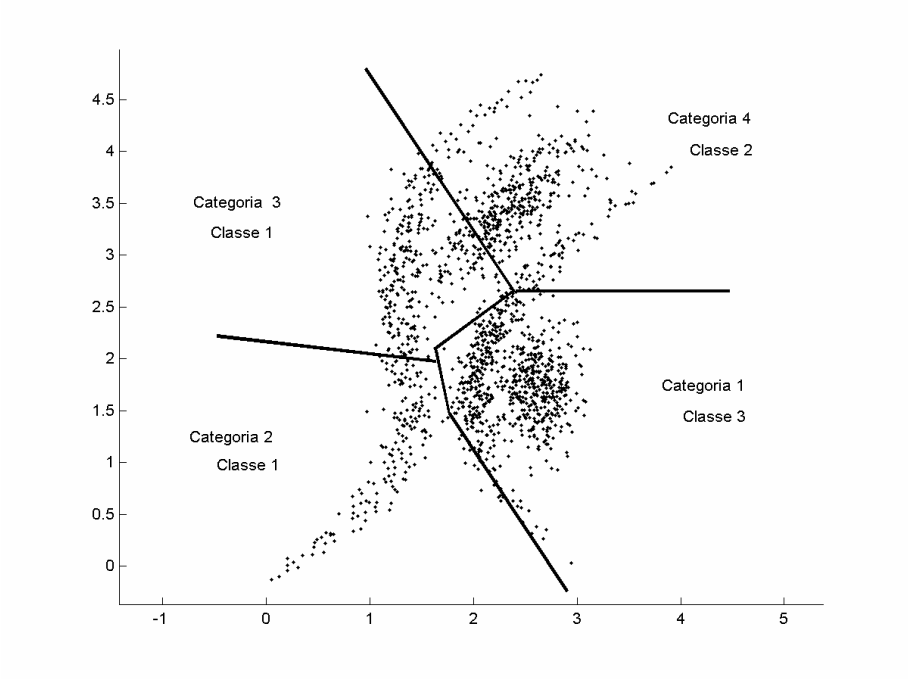

Figura 4.8 - Categorias geradas pela rede neural FKCN na fase de treinamento para os dados de treino, e suas correspondentes classes

As categorias geradas são praticamente idênticas àquelas geradas pelo método *FCM*, o que justifica visualmente as taxas de *acertos rígidos* também quase iguais.

A mesma associação entre categorias e classes pode ser observada tanto para a rede *FKCN* e o método *FCM*. As categorias geradas também têm formato circular. A *matriz de confusão* para os dados de teste podem ser observadas na Tabela 4.7.

Classes	**FKCN**			
	1	**2**	**3**	**4**
1	84	16	0	0
2	21	79	0	0
3	0	0	100	0
4	7	19	74	0

Tabela 4.7 – Matriz de confusão para a execução do FKCM para os dados de teste. As linhas correspondem às classes (valor esperado) e as colunas correspondem às amostras classificadas pelo método (valor estimado). As células sombreadas correspondem ao número de amostras que foram classificadas corretamente em suas respectivas classes

A Figura 4.9 mostra a categorização para m = 1,1 gerada pelo método *GG-FCM*. Esta foi a categorização deste método que obteve maior taxa *acertos rígidos* para os dados de teste.

Pode-se observar os formatos indeterminados das categorias geradas são derivados do formato das categorias geradas pelo método *FCM*, dado que o *GG-FCM* foi inicializado com estas categorias. Para o *GG-FCM* as categorias geradas também não correspondem às classes do problema.

A *categoria 1* englobou pontos amostrais das classes 1 e 4, correspondendo à *classe 1.* A categoria 2 englobou pontos de três classes, a 1, 2 e 4. Ela corresponde à *classe 2*.

A *categoria 3* agrupou pontos das classes 3 e 4, representando então a *classe 3.* A categoria 4 só incluiu pontos da classe 1, correspondendo então à *classe 1*.

Como nos métodos *FCM* e *FKCN*, a *classe 4* também não foi representada pelas categorias geradas pelo método *GG-FCM*, enquanto que a *classe 1* foi representada por 2 categorias.

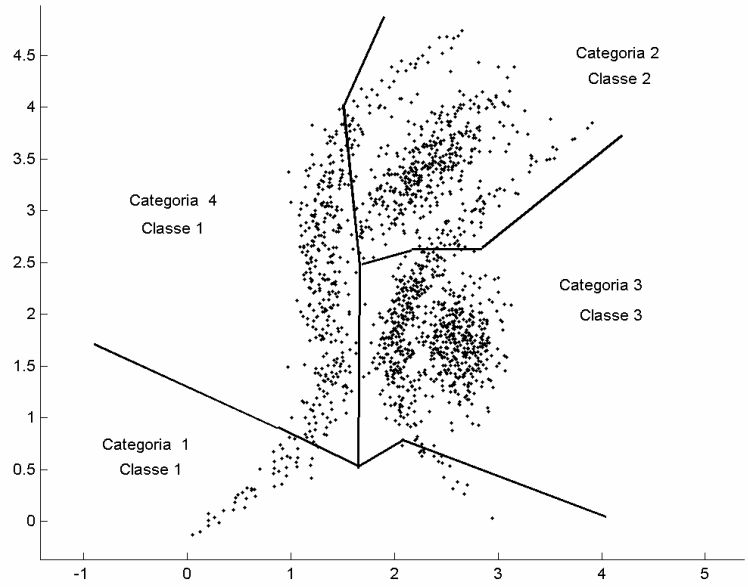

Figura 4.9 – Categorias geradas pelo algoritmo GG-FCM, inicializado com os centros gerados pelo método FCM, para m = 1.1 na fase de treinamento para os dados de treino, e suas correspondentes classes

A Tabela 4.8 mostra a *matriz de confusão* da execução do método *GG-FCM* para os pontos de teste.

Novamente os dados amostrais da *classe 4* foram classificados como pertencentes às outras 3 classes, obtendo taxa de acertos de 0%, e todos os pontos da *classe 3* foram corretamente classificados.

Classes	**GG-FCM**			
	1	**2**	**3**	**4**
1	79	19	2	0
2	3	97	0	0
3	0	0	100	0
4	2	19	79	0

Tabela 4.8 – Matriz de confusão para a execução do GG-FCM para os dados de teste para m = 1,1. As linhas correspondem às classes (valor esperado) e as colunas correspondem as amostras classificadas pelo método (valor estimado). As células sombreadas correspondem ao número de amostras que foram classificadas corretamente em suas respectivas classes

Pela Tabela 4.6 observa-se que grande parte dos pontos classificados como pertencentes à *classe 1* (94,05 %) realmente pertenciam à *classe 1*, enquanto que nenhum ponto foi classificado como pertencente à *classe 4*.

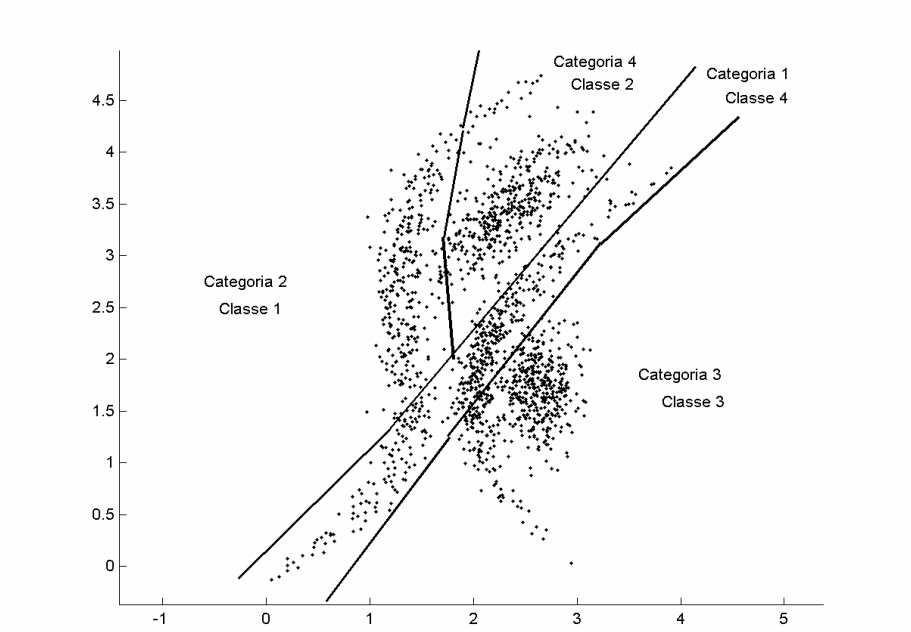

Figura 4.10 – Categorias geradas pelo algoritmo GK para m = 2 na fase de treinamento para os dados de treino, e suas correspondentes classes

A Figura 4.10 mostra a categorização realizada pelo método *GK* para os dados de treino, com $m = 2$, que obteve a melhor taxa *acertos rígidos* dentre os métodos de categorização *FCM*, *FKCN* e *GG-FCM* e *GG*.

Mesmo tendo obtido taxas de *acertos rígidos* altas, as 4 categorias geradas não correspondem exatamente às 4 classes reais do problema.

Porém, diferente das categorizações realizadas pelos métodos *FCM* e *GG* e pela rede neural *FKCN*, o particionamento obtido pelo *GK* conseguiu representar as 4 classes do problema.

Pode-se observar que as categorias geradas têm formato elíptico. Elas também contêm o mesmo número de elementos cada, o que é próprio do método.

A *categoria 1* agrupou pontos das classes 1 e 4, sendo a maioria dos pontos da classe 4. Ela representa então a *classe 4*. A *categoria 2* agrupou somente pontos da *classe 1*, correspondendo a esta classe.

A *categoria 3* englobou pontos das classes 3 e 4, correspondendo à *classe 3* e a *categoria 4* englobou pontos das classes 1 e 2 e corresponde à *classe 2*.

Pela Tabela 4.9, que mostra a matriz de confusão do melhor resultado do método *GK*, observa-se que as classes 2 e 3 obtiveram uma taxa de classificação bem superior às das classes 1 e 4.

Classes	**GK**			
	1	**2**	**3**	**4**
1	77	6	0	17
2	6	94	0	0
3	0	0	97	3
4	0	0	32	68

Tabela 4.9 – Matriz de confusão para a execução do GK para os dados de teste para m = 2. As linhas correspondem às clæsses (valor esperado) e as colunas correspondem as amostras classificadas pelo método (valor estimado). As células sombreadas correspondem ao número de amostras que foram classificadas corretamente em suas respectivas classes

Pela Tabela 4.6 observa-se que a categoria que representa a *classe 1* englobou 92,77% dos pontos da classe 1, enquanto que a categoria que representa a *classe 2* englobou 94% dos pontos da classe 2, indicando uma ótima categorização para estas duas classes.

5 RECONHECIMENTO DE DÍGITOS MANUSCRITOS

O *Reconhecimento de Caracteres Manuscritos* é uma tarefa árdua e complexa, pois é dependente dos diversos tipos de caligrafias, cada qual com suas características pessoais, e da qualidade dos dados originais antes e após a digitalização, dentre outros fatores.

Este capítulo objetiva mostrar a comparação entre o desempenho do sistema proposto *ICC-KNN* e dos demais métodos de categorização *FCM*, *FKCN*, *GG*, *GK* e *K-NN nebuloso* apresentados nesta dissertação quando aplicados ao problema de reconhecimento de dígitos manuscritos.

Os dados amostrais utilizados nesta tese foram obtidos, extraídos e tratados em (RODRIGUES, 2001).

Os dígitos foram extraídos de formulários especialmente preparados e preenchidos por pessoas de diferentes escolaridades que transitavam no campus da UFRJ.

Estes formulários eram compostos de campos quadriculados impressos para delimitar os dados escritos à mão nestes.

Os campos dos formulários foram escaneados e transformados em imagens de formato *Tiff*. A imagem original foi convertida de um *bitmap em escala de cinza* para um *bitmap binário* e foi escalada para caber em uma matriz de 16 por 16 pixels.

A imagem binarizada foi então afinada pelo algoritmo de afinamento apresentado em (PAVLIDIS, 1982)

O afinamento extrai o esqueleto dos dígitos através da preservação de suas características mais relevantes. O esqueleto é obtido através de sucessivas eliminações

de *pixels* que pertencem à borda do dígito e que não segmentam a imagem quando eliminados.

Após o afinamento, a extração de características dos dígitos foi realizada. O método para extrair as características é baseado na projeção do contorno da imagem do dígito nos lados do polígono desenhado ao redor deste.

O vetor de características é então formado pelas distâncias perpendiculares da imagem do dígito aos lados do polígono.

O polígono escolhido nesta dissertação foi o quadrado com 128 características, pois ele obteve um bom resultado de classificação em (RODRIGUES, 2001), próximo ao melhor resultado que foi obtido pelo quadrado com 512 características.

Assim pôde ser mantido o compromisso entre eficiência e precisão, dado que quanto maior o número de características, maior o tempo de execução dos métodos de categorização.

5.1 REPRESENTAÇÃO DOS DADOS

Os dígitos são representados através de vetores de características com 128 coordenadas. Destas 128 características, as coordenadas que tinham valor 0 para todas as amostras foram eliminadas por não adicionarem nenhuma informação relevante ao problema.

Dígitos	Amostras por Classe	Dados de Treino	Dados de Teste
0	665	532	133
1	661	529	132
2	722	578	144
3	378	303	75
4	305	244	61
5	293	235	58
6	278	223	55
7	309	248	61
8	238	191	47
9	228	183	45
Total	4077	3266	811

Tabela 5.1 – Tabela com o número de amostras, os dados de treino e os dados de teste de cada classe e os seus totais

Deste modo, a dimensionalidade do problema foi reduzida para 122 características. Cada dígito corresponde a uma classe do problema. O problema é então formado por 10 classes disjuntas.

O espaço amostral do problema totaliza 4077 dígitos, dos quais 80% de pontos de cada classe formam os dados de treinamento, perfazendo 3266 amostras e 20% formam os dados de testes, perfazendo 811 amostras. A quantidade de amostras por classes é mostrada na Tabela 5.1 acima.

5.2 APLICANDO O SISTEMA ICC-KNN

Esta seção mostra o desempenho do sistema proposto *ICC-KNN* quando aplicado ao problema de reconhecimento de dígitos manuscritos.

Na *primeira fase de treinamento* (*PFT)* do *Módulo de Modelagem*, os dados de treinamento foram divididos em suas respectivas classes e o método *FCM* foi executado para cada classe do problema separadamente, com *c* variando de 2 a 30 categorias para a *constante nebulosa* fixa $m = 1{,}25$.

As melhores partições geradas pelo *FCM* para cada classe foram identificadas pela medida de validação proposta *ICC*. O número de partições que serão os padrões de suas respectivas classes na execução do método *K-NN nebuloso* na *segunda fase de treinamento (SFT)* do *Módulo de Treinamento* do sistema, são mostrados na Tabela 5.2.

Classes	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Partições	22	29	12	25	15	26	25	23	10	30

Tabela 5.2 – Número de partições identificados pela medida de validação ICC para cada classe com 122 características, cujos centros serão usados como padrões no método K-NN nebuloso

Na *SFT*, o método *K-NN nebuloso* foi executado para os dados de treinamento, com os padrões gerados na *PFT*. Para fins de comparação, o *K-NN nebuloso* também foi executado com padrões escolhidos aleatoriamente da massa de dados de treinamento. O número de vizinhos *k* foi variado de 3 a 7 vizinhos para os valores de $m \in \{1{,}1;\ 1{,}25;\ 1{,}5;\ 2\}$.

O gráfico da Figura 5.1 mostra as taxas de *acertos rígidos* obtidas internamente pelo sistema durante a *SFT* durante a escolha dos melhores *k* e *m.*

A *linha sólida* corresponde às taxas de *acertos rígidos* obtidas na *SFT* do sistema *ICC-KNN* para os padrões gerados na *PFT* e a *linha tracejada* corresponde às taxas de *acertos rígidos* obtidas para os padrões aleatórios.

As taxas de *acertos rígidos* para os padrões da *PFT* são maiores para partições mais próximas de partições rígidas, ou seja, quando m = 1,1 e m = 1,25, tendo valores bem próximos.

Para m = 1,5, as taxas de *acertos rígidos* começam a diminuir e para m = 2, elas obtém os menores valores para o problema.

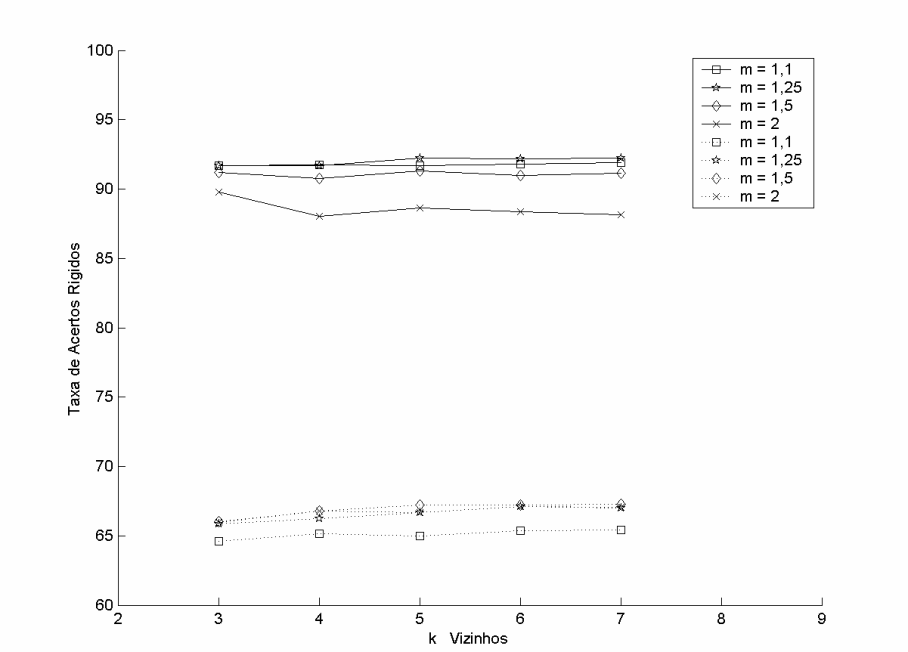

Figura 5.1 – Gráfico da Taxa de Acertos Rígidos em função do número de vizinhos k e da constante nebulosa m. A linha sólida representa os acertos do método K-NN nebuloso para os padrões calculados na primeira fase de treinamento e a linha pontilhada representa os acertos para os padrões escolhidos aleatoriamente

As taxas de *acertos rígidos* para os padrões aleatórios foram melhores para partições mais nebulosas, obtendo os piores valores para m = 1,1. Neste caso, a maior nebulosidade das partições compensou a escolha não otimizada dos padrões.

A maior taxa de *acertos rígidos* para o método *K-NN nebuloso* com os padrões obtidos na *PFT* foi de 92,23%, para $m = 1{,}25$ e $k = 7$, muito superior à obtida para os padrões aleatórios, que foi de 67,27% para $m = 1{,}5$ e $k = 7$.

A taxa de *acertos rígidos* obtida com os padrões da *PFT* foi 37% superior à obtida com os padrões aleatórios.

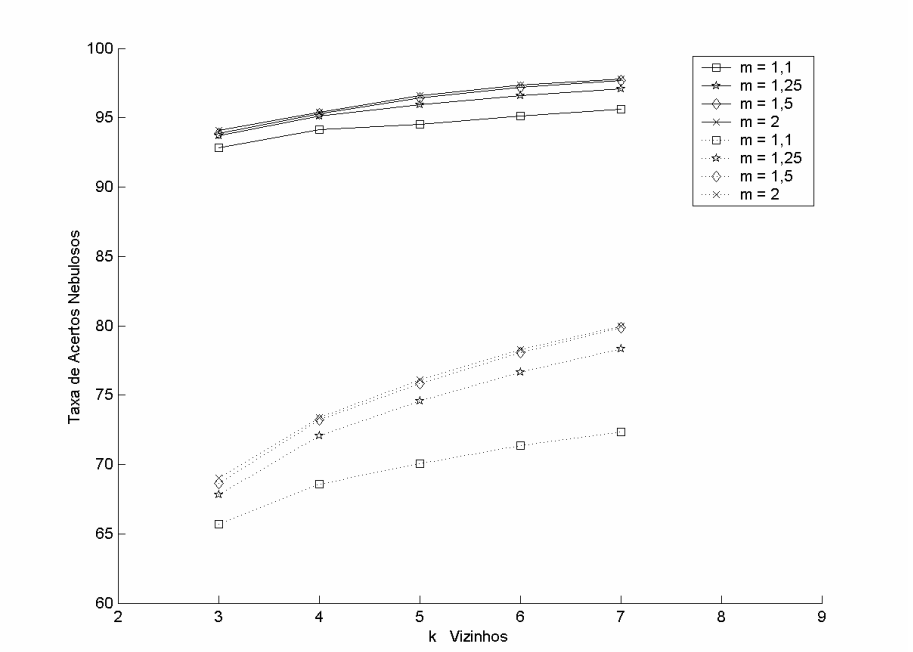

Figura 5.2 – Gráfico da Taxa de Acertos Nebulosos em função do número de vizinhos k e da constante nebulosa m. A linha sólida representa os acertos do método K-NN nebuloso para os padrões da PFT e a linha pontilhada representa os acertos para os padrões aleatórios

As taxas de *acertos nebulosos*, tanto para os padrões da *PFT* como para os padrões aleatórios, são mais dependentes dos valores da constante nebulosa e do número de vizinhos que as taxas de *acertos rígidos*. Isto pode ser observado na Figura 5.2.

Para partições mais nebulosas, as taxas de *acertos nebulosos* são maiores, sendo quase idênticas para $m = 1{,}5$ e $m = 2$, atingindo os piores valores para $m = 1{,}1$ em ambos os casos. Seus valores também aumentam conforme o valor de k aumenta.

As taxas de *acertos rígidos* e as taxas de *acertos nebulosos* obtidas foram maiores para os padrões calculados na *PFT* do sistema *ICC-KNN* do que para os padrões escolhidos aleatoriamente.

Na *SFT* foram encontrados os melhores parâmetros a serem utilizados na execução do método *K-NN nebuloso* para os dados de teste, que são $m = 1,25$ e $k = 7$.

No *Módulo de Reconhecimento de Padrões* do sistema, o método *K-NN nebuloso*, com estes parâmetros, é então aplicado aos dados de teste, tanto para os padrões computados na *PFT* como para os padrões aleatórios.

Para os padrões da *PFT*, a taxa de *acertos rígidos* obtida é de 87,8% e a taxa de *acertos nebulosos* é de 94,53%, ambas maiores que as taxas obtidas para os padrões aleatórios, que foram de 72,4% e 85,63%, respectivamente.

O ganho em utilizar o *K-NN nebuloso* com os padrões da *PFT* em relação aos padrões aleatórios na fase de teste foi de 21,3%. A Tabela 5.3 mostra as taxas de *acertos rígidos* para os métodos e os seus tempos de execução.

Métodos	**ICC-KNN**	**K-NN Neb. Alea.**
Acertos Ríg.	87,8%	72,4%
Acertos Neb.	94,53%	85,63%
Tempos	7166 s	1224,3 s

Tabela 5.3 – Acertos rígidos, acertos nebulosos e tempo de treinamento dos métodos de categorização ICC-KNN e K-NN nebuloso com padrões aleatórios para os dados de treinamento com 122 características, para m = 1,25 e k = 7

Este resultado mostra a validade da utilização o sistema *ICC-KNN* para calcular os padrões e os parâmetros a serem utilizados pelo *K-NN nebuloso* no problema de reconhecimento de dígitos.

A Tabela 5.5 ilustra as *matrizes de confusão* para os resultados descritos acima e a Tabela 5.4 mostra dentre todos os pontos que foram classificados como pertencentes a cada classe, a percentagem de pontos que realmente pertencem a esta classe e foram classificados como tais.

Classe	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
PFT	95,4%	85,8%	94,7%	88,9%	96,5%	91,7%	94,2%	75,8%	61%	80%
Aleat.	81%	84,2%	90,4%	72,7%	73,4%	74,6%	69,8%	57,5%	36,4%	55,4%

Tabela 5.4 - Percentagem dos acertos rígidos para cada classe do problema, geradas pelo K-NN nebuloso aplicado aos dados de teste para os padrões da PFT e para os escolhidos aleatoriamente, com m = 1,25 e k = 7

Observa-se que para os padrões calculados na *PFT*, a classe do dígito 4 é a que agrupou mais dados da sua própria classe, enquanto que, para os padrões aleatórios, a classe 2 é a vencedora.

Número de Amostras	Classes	Padrões da PFT									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
133	**0**	124	2	1	0	0	0	1	0	5	0
132	**1**	0	115	2	1	1	0	0	6	2	5
144	**2**	1	5	124	4	0	1	0	2	6	1
75	**3**	0	1	1	64	0	0	0	5	3	1
61	**4**	0	3	0	0	55	0	0	1	0	2
58	**5**	0	0	1	2	0	55	0	0	0	0
55	**6**	1	0	2	0	0	0	49	1	2	0
61	**7**	0	6	0	0	0	0	0	50	4	1
47	**8**	4	2	0	1	0	1	2	1	36	0
45	**9**	0	0	0	0	1	3	0	0	1	40

Tabela 5.5 – Matriz de confusão para a execução do K-NN nebuloso para os dados de teste, com m = 1,25 e k = 7, para os padrões da PFT e para os escolhidos aleatoriamente. As linhas correspondem às classes e as colunas correspondem às amostras classificadas pelo método. As células sombreadas correspondem ao número de amostras que foram classificadas corretamente em suas respectivas classes

O tempo de computação para a *PFT*, para os dados com 122 características, foi de 6099 segundos. Nesta fase, o método *FCM* e a medida de validação *ICC* foram executados para cada classe separadamente. A Tabela 5.6 mostra os tempos de execução do método *FCM* para cada classe.

	T0	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**Total**
FCM	965s	1055s	1257s	454s	261s	560s	411s	366s	380s	381s	6090s

Tabela 5.6 – Tempos de execução do método FCM para cada classe, de 0 a 9, e o Tempo Total de execução para todas as classes, para 122 características

A *SFT*, onde são variados os valores de *m* e de *k* na execução do método *K-NN nebuloso* e são calculadas as taxas de *acertos rígidos*, durou 1076 segundos. O *tempo total de treinamento* do sistema *ICC-KNN* foi de 7166 segundos, ou seja, de 1h e 59 seg.

O tempo de execução do *K-NN nebuloso* para com os padrões aleatórios é formado pelo tempo da escolha aleatória dos padrões e pelo tempo da *SFT*, dado que a *PFT* não faz parte de sua execução.

O tempo da escolha dos padrões é pequeno, sendo de 0,3 segundo e o tempo da *SFT* é de 1224 segundos, perfazendo o tempo total de execução de 1224 segundos. Seu tempo de execução é inferior ao tempo do *K-NN nebuloso* com os padrões escolhidos na *PFT*, porém suas taxas de acertos também são inferiores.

5.2.1 COMPARANDO O SISTEMA ICC-KNN COM OS MÉTODOS DE CATEGORIZAÇÃO ESTUDADOS

Os métodos de categorização *FCM*, *FKCN*, *GG* e *GK* também foram aplicados como componentes de *Sistemas de Reconhecimento Estatístico de Padrões*, conforme explanado na seção 4.4, ao problema de reconhecimento de dígitos a fim de avaliar seu desempenho juntamente com o desempenho do sistema *ICC-KNN*.

Porém os métodos *GG* e *GK*, ao contrário dos métodos *FCM*, *FKCN* e do próprio *ICC-KNN*, oferecem restrições (seções 2.2.2 e 2.2.3) ao conjunto de amostras a serem categorizadas, pois envolve o calculo de matrizes inversas.

Para aplicar estes métodos às amostras, foi necessário então reduzir o número de características destas. A técnica utilizada com este intuito foi a *Análise dos Componentes Principais – PCA* (Principal Components Analysis). (KUPAC, 2000)

O *PCA* reduz a dimensão do espaço vetorial através de combinações lineares das variáveis principais, ou seja, das variáveis que oferecem maior variância ao conjunto de dados.

Estas combinações geram um novo sistema de coordenadas, através da rotação do sistema original, que preserva a variância total do conjunto de amostras e o máximo possível de informações relevantes.

A classe com o menor número de amostras é a classe do dígito 9, com 183 amostras para a fase de treinamento. Então para aplicar os métodos *GG* e *GK* a este problema, o número de características p deve obedecer à equação $p(p-1)/2 < 183$, de modo que $p = 19$ deve ser o novo número de características das amostras. (seção 2.2.2)

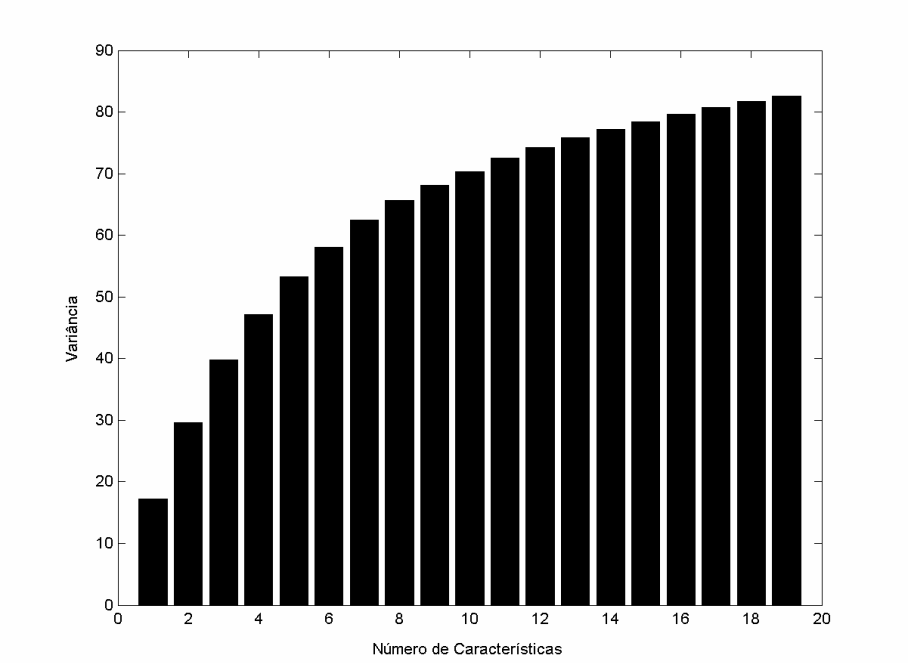

Figura 5.3 – Gráfico da Variância em função do número de características das amostras dos dígitos manuscritos

O *PCA* foi aplicado aos dados, originalmente com 122 características. Ao reduzir para 19 as características das amostras, a fração da variância original dos dados preservada foi de 82,6%. O gráfico da variância em função do número de características para os dados deste problema é mostrado na Figura 5.3.

Para fins comparativos, o sistema *ICC-KNN* também foi executado para os dados transformados com 19 características, juntamente com os métodos de categorização *FCM*, *FKCN*, *GG*, *GK*, para os mesmos conjuntos de treinamento e teste.

Na *PFT* do sistema *ICC-KNN*, a medida de validação *ICC* encontrou as melhores partições geradas pelo método *FCM* com os dados de treinamento, que são mostradas na Tabela 5.7.

Classes	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Partições	24	28	12	24	10	28	23	25	13	28

Tabela 5.7 – Número de partições identificados pela medida de validação ICC para cada classe com 19 características, cujos centros serão usados como padrões no método K-NN nebuloso

A *SFT* foi executada para os padrões calculados na *PFT* e para os padrões aleatórios. Para os padrões da *PFT*, a maior taxa de *acertos rígidos* foi de 92% para m = 1,25 e k = 6 e a maior taxa de *acertos nebulosos* foi de 97,8% para m = 2 e k = 7.

Para os padrões escolhidos aleatoriamente, a maior taxa de *acertos rígidos* foi de 71,4% para m = 2 e k = 6 e a maior taxa de *acertos nebulosos* foi de 82% para os valores de m = 2 e k = 7.

O *K-NN nebuloso* foi então executado no *Módulo de Reconhecimento de Padrões* com os dados de teste, para os parâmetros que obtiveram as maiores taxas de *acertos rígidos* na *SFT*, que foram m = 1,25 e k = 6, para os padrões da *PFT* e para os aleatórios.

Para os padrões da *PFT*, a taxa de *acertos rígidos* obtida é de 86,7% e a taxa de *acertos nebulosos* é de 93,8%, ambas maiores que as taxas obtidas para os padrões aleatórios, que foram de 75,22% e 85,66%, respectivamente.

Observa-se que as taxas de *acertos rígidos e nebulosos* obtidas com os padrões da *PFT* são maiores para as amostras com 122 características (87,8% e 94,53%, respectivamente) do que para as amostras com 19 características.

Já para os padrões aleatórios, a taxa de *acertos rígidos* é maior para amostras com 19 características e a taxa de *acertos nebulosos* é praticamente igual para ambos os casos. Para 122 características, as taxas obtidas foram de 72,4% e 85,63%, respectivamente.

O tempo total de execução do sistema *ICC-KNN* foi de 1784 segundos, onde o tempo de execução da *PFT* foi de 1569s e o da *SFT* foi de 215s. Os tempos de execução do método *FCM* para cada classe pode ser comparado na Tabela 5.8.

Para os padrões aleatórios, o tempo total de execução foi de 260 segundos.

	T0	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**Total**
FCM	271s	211s	280s	123s	122s	94s	97s	116s	107s	144s	1565s

Tabela 5.8 – Tempos de execução do método FCM para cada classe, de 0 a 9, e o Tempo Total de execução para todas as classes, para 19 características

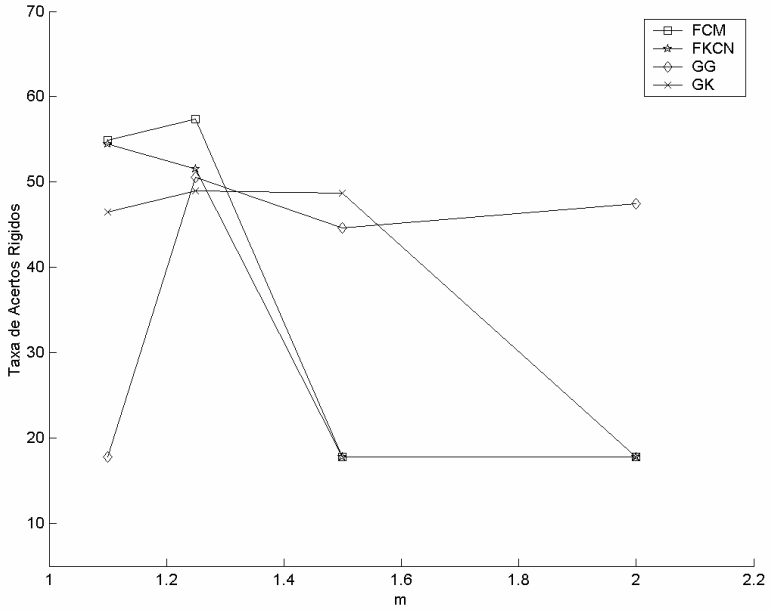

Figura 5.4 – Gráfico da Taxa de Acertos Rígidos dos métodos FCM, FKCN, GG e GK em função da constante nebulosa m para os dados de teste.

Os métodos de categorização *FCM*, *FKCN*, *GG* e *GK* foram executados para c = 10 categorias, dado que se sabe que o problema tem 10 classes, e para os valores da constante nebulosa $m \in \{1,1;\ 1,25\ ;\ 1,5;\ 2\}$.

As taxas de *acertos rígidos* e de *acertos nebulosos* para os métodos de categorizações acima podem ser observados na Figura 5.4 e na Figura 5.5. Pode-se observar que os resultados são altamente dependentes dos valores da *constante nebulosa m*.

Métodos	ICC-KNN	K-NN pad. alea.	FCM	FKCN	GG	GK
Acertos Ríg.	86,7%	75,22%	57%	55%	51%	49%
Acertos Neb.	93,8%	85,66%	60%	54%	39,5%	39,8%
Tempos	1784 s	260 s	30,38 s	32,79 s	108,15 s	711,77 s

Tabela 5.9 – Acertos rígidos, acertos nebulosos e tempo de treinamento dos métodos de categorização ICC-KNN , K-NN nebuloso com padrões aleatórios, FCM, FKCN, GG e GK para os dados de teste com 19 características

O melhor desempenho dentre estes métodos foi do método *FCM*, com taxa de *acertos rígidos* de 57% e taxa de *acertos nebulosos* de 60%, ambos para a constante nebulosa m = 1,25.

Porém seu desempenho foi muito inferior ao obtido pelo sistema *ICC-KNN*. O ganho rígido do sistema *ICC-KNN* sobre o método *FCM* é de 52%. A Tabela 5.9 mostra as taxas de acertos e os tempos de execução dos métodos para as amostras de teste com 19 características.

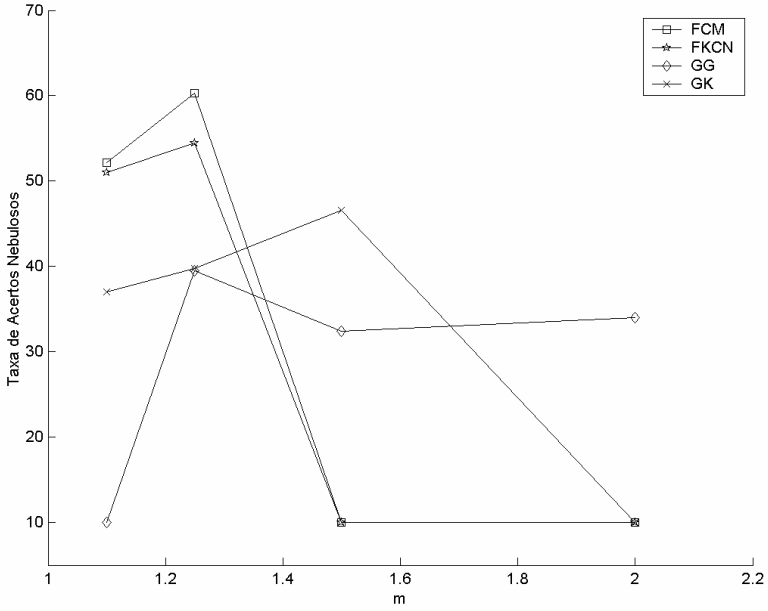

Figura 5.5 - Gráfico da Taxa de Acertos Nebulosos dos métodos FCM, FKCN, GG e GK em função da constante nebulosa m para os dados de teste.

O tempo de execução do sistema *ICC-KNN* é maior que o tempo dos outros métodos de categorização, sendo compensado pelo seu maior poder de classificação.

As execuções do método *FCM* e da rede *FKCN* para $m = 1{,}5$ e $m = 2$ geraram categorias cujos centros convergiram para o mesmo ponto. Assim todos os dígitos foram classificados como pertencentes a uma única classe.

Para $m = 1{,}1$ e $m = 1{,}25$, quando as partições são mais nebulosas, as taxas de *acertos rígidos* do *FCM* foram de 55% e 57%, enquanto que as da rede *FKCN* foram de 55% e 52%, respectivamente.

As taxas de *acertos nebulosos* também foram baixas, sendo para o *FCM* de 52% e 60% e para a rede *FKCN* de 51% e 54%, para $m = 1{,}1$ e $m = 1{,}25$ respectivamente.

Número de Amostras	Classes	Padrões da PFT									
		0	1	2	3	4	5	6	7	8	9
133	0	112	7	10	0	0	0	1	3	0	0
132	1	0	83	15	1	5	0	3	25	0	0
144	2	7	8	106	6	1	0	12	4	0	0
75	3	0	7	12	36	0	0	1	19	0	0
61	4	0	3	1	0	45	0	2	10	0	0
58	5	0	3	20	24	1	0	5	5	0	0
55	6	3	6	9	0	0	0	37	0	0	0
61	7	0	4	11	0	0	0	0	46	0	0
47	8	11	7	15	0	0	0	5	9	0	0
45	9	1	0	4	5	0	0	0	35	0	0

Tabela 5.10 – Matriz de confusão para a execução do método FCM para os dados de teste, com m = 1,25. As linhas correspondem às classes e as colunas correspondem às amostras classificadas pelo método. As células sombreadas correspondem ao número de amostras que foram classificadas corretamente em suas respectivas classes

A Tabela 5.10 e a Tabela 5.11 ilustram as *matrizes de confusão* para os melhores resultados obtidos pelo método *FCM* e a rede *FKCN*, respectivamente.

Número de Amostras	Classes	Padrões da PFT									
		0	1	2	3	4	5	6	7	8	9
133	0	115	12	1	0	0	1	1	3	0	0
132	1	0	103	3	0	19	1	6	0	0	0
144	2	13	23	74	3	1	7	20	3	0	0
75	3	0	16	12	37	4	1	0	5	0	0
61	4	0	7	0	0	36	3	7	8	0	0
58	5	0	3	5	27	1	18	2	2	0	0
55	6	4	11	1	0	0	2	37	0	0	0
61	7	2	13	2	0	20	2	0	22	0	0
47	8	14	20	1	1	4	3	4	0	0	0
45	9	1	3	0	9	24	3	0	5	0	0

Tabela 5.11 – Matriz de confusão para a execução da rede neural FKCN para os dados de teste, com m = 1,1. As linhas correspondem às classes e as colunas correspondem às amostras classificadas pelo método. As células sombreadas correspondem ao número de amostras que foram classificadas corretamente em suas respectivas classes

Observa-se que nos dois casos, as categorias dos dígitos 8 e 9 não foram representadas por nenhuma categoria, obtendo taxa de acertos de 0%. Estes dígitos foram classificados como pertencentes às demais classes.

A Tabela 5.12 mostra a percentagem de pontos que pertencem a cada classe e foram classificados como tais. Tanto para o método *FCM* como para a rede *FKCN*, a

categoria que representa a *classe 0* foi a que englobou mais pontos pertencentes à sua própria classe.

Classe	0	1	2	3	4	5	6	7	8	9
FCM	83,6%	64,8%	52,2%	50%	86,5%	0%	56%	29,5%	0%	0%
FKCN	77,2%	49%	74,8%	48%	33%	44%	48%	46%	0%	0%
GG	100%	88,9%	43,3%	54,5%	24,2%	48%	0%	26%	0%	0%
GK	84,2%	80,3%	85,4%	0%	70,5%	36,2%	0%	41%	0%	0%

Tabela 5.12 - Percentagem de pontos que pertencem a cada classe e foram classificados como tais, em relação ao total de pontos classificados como pertencentes a cada classe, para os métodos FCM, FKCN, GG e GK.

O método *GG* foi inicializado pelos centros gerados pelo método *FCM*. Suas execuções, ao contrário das dos métodos *FCM* e *FKCN*, geraram centros coincidentes para *m = 1,1*, ou seja, para as partições mais rígidas.

Sua maior taxa de *acertos rígidos* foi de 51% e a maior taxa de *acertos nebulosos* foi de 39,5%, ambas para *m = 1,25* e ambas inferiores às obtidas pelo método *FCM* e pela rede *FKCN*.

Número de Amostras	Classes	Padrões da PFT									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
133	**0**	92	0	36	1	0	2	0	2	0	0
132	**1**	0	80	26	0	23	0	0	3	0	0
144	**2**	0	1	126	0	5	8	0	4	0	0
75	**3**	0	2	17	30	14	2	0	10	0	0
61	**4**	0	3	2	0	29	0	0	27	0	0
58	**5**	0	0	5	21	5	23	0	4	0	0
55	**6**	0	0	49	0	1	5	0	0	0	0
61	**7**	0	2	6	0	21	2	0	30	0	0
47	**8**	0	0	23	0	13	4	0	7	0	0
45	**9**	0	2	1	3	9	2	0	28	0	0

Tabela 5.13 – Matriz de confusão para a execução do método GG inicializado com os centros gerados pelo método FCM para os dados de teste, com m = 1,25. As linhas correspondem às classes e as colunas correspondem às amostras classificadas pelo método. As células sombreadas correspondem ao número de amostras que foram classificadas corretamente em suas respectivas classes

A Tabela 5.13 mostra a matriz de confusão do melhor resultado obtido pelo método *GG*. As classes dos dígitos 6, 8 e 9 não foram representadas por nenhuma categoria, obtendo 0% de acertos.

Pela Tabela 5.12 observa-se que 100% dos pontos classificados como pertencentes à categoria do digito 0 eram realmente desta classe, sendo esta uma classe otimamente categorizada.

O método *GK* obteve as piores taxas de acertos rígidos, de 49% para *m = 1,25*. Nas execuções do método para *m = 2*, para as partições mais nebulosas, todos os centros gerados convergiram para o mesmo ponto.

Número de Amostras	Classes	Padrões da PFT									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
133	**0**	108	3	3	0	18	1	0	0	0	0
132	**1**	0	105	12	0	14	1	0	0	0	0
144	**2**	0	6	106	0	8	12	0	12	0	0
75	**3**	0	15	39	0	1	3	0	17	0	0
61	**4**	0	10	9	0	42	0	0	0	0	0
58	**5**	0	5	26	0	10	17	0	0	0	0
55	**6**	0	8	18	0	29	0	0	0	0	0
61	**7**	1	13	21	0	6	1	0	19	0	0
47	**8**	2	10	4	0	23	3	0	5	0	0
45	**9**	0	7	8	0	23	2	0	5	0	0

Tabela 5.14 – Matriz de confusão para a execução do método GK para os dados de teste, com m = 1,25. As linhas correspondem às classes e as colunas correspondem às amostras classificadas pelo método. As células sombreadas correspondem ao número de amostras que foram classificadas corretamente em suas respectivas classes

A Tabela 5.14 mostra a matriz de confusão para o melhor resultado obtido pelo método *GK*. As classes dos dígitos 3, 6, 8 e 9 não foram representadas por nenhuma categoria.

A Tabela 5.12 mostra que foi a categoria do dígito 2 que mais agrupou pontos dae sua própria classe, sendo 85,4% de seus pontos dígitos 2.

Os resultados dos métodos de categorização *FCM*, *FKCN*, *GG* e *GK* foram inferiores aos obtidos pelo sistema *ICC-KNN*. Isto demonstra a maior versatilidade do sistema em lidar com dados mais complexos e com categorias com alta sobreposição.

6 CONCLUSÕES

O objetivo deste trabalho de pesquisa foi estudar os diversos métodos de categorização nebulosos existentes na literatura e suas aplicações em problemas de reconhecimento de padrões.

Através deste estudo, observou-se que uma dificuldade enfrentada pelos métodos era encontrar o número ideal de partições que melhor representasse um conjunto de amostras.

Existem várias métricas nebulosas utilizadas para avaliar o número ideal de partições, porém cada uma analisa um determinado aspecto do conjunto amostral.

Observou-se a possibilidade de criar uma nova medida de validação que pudesse analisar tanto espaços rígidos como nebulosos levando em conta a maior gama de aspectos do espaço amostral. A *ICC* foi então cunhada empiricamente em decorrência da análise das funcionalidades e restrições das métricas nebulosas estudadas na seção 2.3 e do FLD.

Os métodos de categorização nebulosos não podem ser utilizados independentemente com a finalidade de reconhecer padrões, dado que eles não fornecem meios de associar as categorias geradas com as classes do problema.

Porém, associados, eles demonstraram um grande potencial classificador. A análise e avaliação deste potencial culminaram na criação de um Sistema de Reconhecimento Estatístico de Padrões, o *Sistema ICC-KNN*.

6.1 RESULTADOS OBTIDOS

6.1.1 O EFLD

O *FLD* é uma medida utilizada largamente para avaliar partições geradas por métodos de categorização rígidos.

Sua versão estendida definida, o *EFLD*[5], estendeu eficientemente as funcionalidades do *FLD*, tornando possível utilizar esta medida para avaliar partições nebulosas além das rígidas.

Além disto, a otimização feita culminou na sua maior velocidade de cálculo.

6.1.2 MEDIDA DE VALIDAÇÃO ICC

A medida de validação *ICC* mostrou ser, nos experimentos realizados nesta pesquisa, uma medida eficiente ao avaliar as partições de espaços rígidos e nebulosos, mesmo nos casos mais difíceis, quando a sobreposição das classes é alta.

Comparada às outras métricas utilizadas com o mesmo intuito, ela conseguiu avaliar os dois aspectos mais importantes de um conjunto de partições: a compacidade das partições e a separação entre elas.

Além disto, ela provou ser uma medida com alto grau de acertos e rápida, obtendo ótimos tempos de execução em relação às outras medidas de validação.

6.1.3 SISTEMA ICC-KNN

O *Sistema ICC-KNN* é um sistema de *Reconhecimento Estatístico Não-Paramétrico de Reconhecimento de Padrões* que uniu as vantagens de dois métodos de categorizações nebulosos: o *FCM* e o *K-NN nebuloso*.

O método *FCM* é aplicado para particionar o conjunto amostral, enquanto a medida de validação proposta *ICC* avalia o melhor número de categorias gerado.

[5] Esta medida foi definida neste trabalho

O método *K-NN nebuloso* funciona então como o reconhecedor de padrões, que associa as novas amostras às categorias geradas.

O sistema obteve uma melhor classificação dos dados, mostrando maior eficiência em relação aos sistemas que utilizam outros métodos de categorização nebulosos.

A vantagem deste sistema é sua facilidade de implementação, dado que foram utilizados dois métodos vastamente conhecidos e que os seus mecanismos de funcionamento não foram alterados.

Outra vantagem é que o sistema não oferece restrição ao conjunto de amostras analisado, podendo ser utilizado em qualquer problema de reconhecimento de padrões.

4.2.1.1 RECONHECIMENTO DE DÍGITOS MANUSCRITOS

O *Sistema ICC-KNN* foi aplicado ao problema de reconhecimento de dígitos manuscritos.

O *ICC-KNN* confirmou novamente sua eficiência através de suas taxas de acertos de classificação que foram superiores as obtidas por sistemas que utilizaram métodos *FCM*, *GG*, *GK* e *FKCN* como categorizadores.

6.2 DIFICULDADES ENCONTRADAS

O maior impasse enfrentado na condução deste trabalho foi a dificuldade em encontrar artigos e livros que contivessem as definições dos métodos de categorização nebulosos e das medidas de validação.

O *FCM* é o método mais aplicado, por ser o mais antigo, rápido e fácil de implementar. Este método é descrito em muitos artigos e livros.

Porém, os métodos *GG* e *GK* são descritos praticamente apenas em seus artigos de origem, que não estão disponíveis via Internet. Estes métodos são pouco empregados na literatura, provavelmente por imporem restrições ao conjunto de amostras e ou por exigirem uma implementação mais complexa.

O mesmo acontece com as medidas de validação, que são apenas referenciadas nos artigos que as utilizam enquanto que seus artigos de origem não estão disponíveis na Internet.

Outra dificuldade enfrentada foi a obtenção de dados para testar os algoritmos, sendo necessário criar dados aleatórios.

6.3 TRABALHOS FUTUROS

Em futuras implementações, o sistema *ICC-KNN* pode ser alterado através da substituição do método *FCM* por outros métodos de categorização existentes. Deste modo, pode ser avaliada a capacidade de classificação do sistema para os diversos critérios de minimização.

Nesta mesma linha, pode-se variar os valores da *constante nebulosa* na execução dos diferentes métodos a fim de encontrar a que obtém melhores resultados.

O sistema *ICC-KNN* fornece como saída os *graus de inclusão* das amostras nas categorias em que foram classificadas, dado que estamos referenciando um sistema nebuloso.

Uma futura implementação pode avaliar estes *graus de inclusão* empregando uma rede neural, como uma *MLP*, criando assim um Sistema Híbrido derivado do *ICC-KNN*.

A partir da classificação obtida, a rede neural pode avaliar as amostras em um espaço dimensional menor, diminuindo seu tempo de treinamento e avaliando somente as classes que realmente são relevantes para o problema.

Outro futuro trabalho é a extensão da medida de validação *ICC* a fim de avaliar casos em que não é obedecida a condição do somatório dos graus de inclusão de cada ponto em todas as categorias ser igual a 1.

7 REFERÊNCIAS BIBLIOGRÁFICAS

ALSABTI, K., RANKA, S., SINGH, V. **An Efficient K-Means Clustering Algorithm**. In: IPPS: 11th International Parallel Processing Symposium, 1998. Disponível na INTERNET via ftp.cise.ufl.edu/pub/faculty/ranka/Proceedings/p8.ps. Arquivo consultado em 1999.

BEICHEL, R., BOLTER, R., PINZ, A. **Fuzzy Clustering of a Landsat TM Scene**. Disponível na INTERNET via http://www.icg.tu-graz.ac.at/~bolter/Publications/igarss99/C07_10-50.pdf Arquivo consultado em 1999.

BEZDEK, J., GRIMBALL, N., CARSON, J., ROSS, T., Structural Failure Determination with Fuzzy Sets, **Civil Engeneering Systems**, v. 3, p. 82-92, 1986

BISHOP, C.M. **Neural Networks for Pattern Recognition**. Oxford University Press, 1995.

BRAGA, Antônio de Pádua, CARVALHO, André Ponce de Leon F. de, LUDERMIR, Teresa Bernarda. Fundamentos de Redes Neurais Artificiais. In: 11ª. ESCOLA DE COMPUTAÇÃO, Rio de Janeiro, Jul. 1998

CANNON, R.L., DAVE, J.V., BEZDEK, J.C. Efficient Implementation of the Fuzzy c-Means Clustering Algorithms. **IEEE Transactions on Pattern Analysis and Machine Intelligence.** v. 8, n.2, p. 248-255, 1986.

DASH, M., LIU, H., YAO, J. **Dimensionality Reduction for Unsupervised Data**. Disponível na INTERNET via www.comp.nus.edu.sg/~liuh/tai97.ps Arquivo consultado em 2000.

DUCH, W., GRUDZINSKI, K. "The weighted k-NN with selection of features and its neural realization". In: Fourth Conference on Neural Networks and Their Applications, Zakopane, p. 191-196, May 1999.

DUDA, R.O., HART, P.E. **Pattern Classification and Scene Analysis**. John Wiley & Sons, 1973. p. 114-121.

FASULO, D. **An Analysis of Recent Work on Clustering Search in High Dimensions.** Disponível na INTERNET via simon.cs.cornell.edu/home/kleinber/stoc97-nn.ps Arquivo consultado em 2000.

FRANCO, C.R., VIDAL, L.S., CRUZ, A.J.O. "A Validity Measure for Hard and Fuzzy Clustering derived from Fisher's Linear Discriminant". In: 2002 IEEE INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS (FUZZ-IEEE2002) as

part of the WORLD CONGRESS ON COMPUTATIONAL INTELLIGENCE (WCCI), Maio, 2002. Honolulu. *A ser publicado em maio de 2002.*

GATH, I., GEVA A.B. Unsupervised optimal fuzzy clustering. **IEEE Transaction on Pattern Analysis and Machine Intelligence**, v. 11, n. 7, p.773-781, July 1989.

GORDON, H.L., SOMORJAI, R.L. Fuzzy Cluster Analysis of Molecular Dynamics Trajectories, **PROTEINS: Structure, Function and Genetics**, v. 14, p. 249-264, 1992.

GHOSH, D., SHIVAPRASAD, A.P. **Possibilistic Clustering in Kohonen Networks for Vector Quantization**. Dep. of Electrical Communication – Indian Institute of Science. Disponível na INTERNET via www.computer.org/students/looking/spring97/ghosh/ Arquivo consultado em 2001

GUSTAFSON, D.E., KESSEL, W.C. Fuzzy clustering with a fuzzy covariance matrix. In: IEEE CONFERENCE ON DECISION AND CONTROL, Jan. 1979. **Proceedings**. p. 761-766.

HAN, J.H., KIM, Y.K. "A Fuzzy K-NN Algorithm Using Weights from the Variance of Membership Values". In: PROCEEDINGS OF THE COMPUTER VISION AND PATTERN RECOGNITION, v. 2, 1998. Disponível na INTERNET via http://computer.org/proceedings/cvpr/0149//volume2/01492394abs.htm. Arquivo consultado em 2000.

HOPPNER, F., KLAWONN, F. **Fuzzy Clustering of Sampled Functions.** Disponível na INTERNET via www.et-inf.fho-emden.de/~dmlab/fc/paper/Hoeppner-NAFIPS-2000.ps.gz Arquivo consultado em 2001

HORIKAWA, S. Fuzzy Classification System using Self-Organizing Feature Map. **Oki Technical Review**, v. 63, n. 159, July 1997.

JDIWIK, A., CHMIELIWSKY, L. CUIDNY, W. and SKTODOWSKI, M. "A 1-NN Preclassifier for Fuzzy k-NN Rule". In: INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION (ICPR '96), v. 4, August 25-29, 1996.

KELLER, J.M., GRAY, M.R., GIVENS Jr., J.A. A Fuzzy K-Nearest Neighbor Algorithm. **IEEE Transaction on Systems, Man and Cybernetics**, v. SMC-15, n. 4, p. 580-585, July/August 1985.

KLAWONN, F., KRUSE, R., TIMM, H. **Fuzzy Shell Cluster Analysis**. Disponível na INTERNET via ftp.et-inf.fho-emden.de/pub/FHO/Informatik/klawonn/heiko.ps.gz Arquivo consultado em 2001.

KLEINBERG, J.M., **Two Algorithms for Nearest-Neighbor search in High Dimensions.** Disponível na INTERNET via simon.cs.cornell.edu/home/kleinber/stoc97-nn.ps Artigo consultado em 2001.

KOSANOVIC, B.R. **Signal and System Analysis in Fuzzy Information Space**. PhD Thesis, University of Pittsburgh, 1995a. Disponível na INTERNET via www.neuronet.pitt.edu/~bogdan/ Arquivo consultado em 2001

_____. **FCMeans Clustering MATLAB Toolbox V2-0**. 1995b. Disponível na INTERNET via www.neuronet.pitt.edu/~bogdan/ Arquivo consultado em 2001. *Apud* BEZDEK, J.C. **Pattern Recognition with Fuzzy Objective Function Algorithms**. Plenum Press, New York, 1981.

KUPAC, G.V. **Sistemas Híbridos Inteligentes para Reconhecimento de Padroes.** Dissertação (Mestrado), Universidade Federal do Rio de Janeiro, IM/NCE, 2000

LAVINE. B.K., **Clustering and Classification of Analytical Data**. Disponível na INTERNET via www.wiley.com/wileychi/eac/pdf/A5204-W.PDF. Arquivo consultado em 1999.

LOONEY, C. N **A Fuzzy Clustering and Fuzzy Merging Algorithm**. Dsponível na INTERNET via http://pinon.cs.unr.edu/~looney/cs479/cs4795.htm Arquivo consultado em 2000.

MARI, M., DELLEPIANE, S. "A Segmentation Method based on Fuzzy Topology and Clustering". In: INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION (ICPR '96), v. 2, August 25-29, 1996.

PAL, N.R., BEZDEK J.C. On Cluster Validity for the Fuzzy c-Means Model. **IEEE Transactions on Fuzzy Systems**, v. 3, n. 3, August 1995.

PAVLIDIS, T. **Algorithms for Graphics and Image Processing.** Computer Science Press, Inc., 1982

_____. **Exploring Gene Expression Data with Class Scores**. To appear in Proceedings of the Pacific Symposium on Biocomputing, 2001. Disponível na INTERNET via www.cs.columbia.edu/~bgrundy/papers/ermgene.pdf Arquivo consultado em 2001.

RIVERA, F.F., ZAPATA, E.L. Cluster validity based on the hard tendency of the fuzzy classification. **Pattern Recognition Letter**, v. 11, p. 7-12, January 1990.

RODRIGUES, R.J., SILVA, E., THOMÉ, A.C.G. Feature Extraction Using Contour Projection In: THE 5TH WORLD MULTI-CONFERENCE ON SYSTEMICS, July 2001. **Proceedings.** Florida

ROSS, T.J. **Fuzzy logic with engineering applications**. McGraw-Hill International Editions – Electrical Engineering Series, 1997.

ROUBENS, M. Pattern Classification Problems and Fuzzy Sets. **Fuzzy Sets and Systems**, v. 1, p. 239-252, 1978.

_____. Fuzzy Clustering Algorithms and their Cluster Validity. **European Journal of Operational Research**, v.10, p. 294-301, 1982.

SANCHEZ, J.S., PLA, F., FERRI, F.J. "Learning Vector Quantization With Alternative Distance Criteria". In PROCEEDINGS OF THE 10TH INTERNATIONAL CONFERENCE ON IMAGE ANALYSIS AND PROCESSING, 1998. Disponível na INTERNET via computer.org/Proceedings/iciap/0040/00400084abs.htm. Arquivo consultado em 2000.

VIVARELLI, F., WILLIAMS, C.K. **Discovering hidden features with Gaussian processes regression** In: Advances in Neural Information Processing Systems 11, eds. M. J. Kearns, S. A. Solla and D. A. Cohn. MIT Press, 1999. Disponível na INTERNET via www.dai.ed.ac.uk/homes/ckiw/./online_pubs.html. Arquivo consultado em 2001

WAKAMI, N., NOMURA, H., ARAKI, S. Fuzzy Logic for Home Alliances. In CHEN, C. H., **Fuzzy Logic and Neural Network Handbook**, McGraw-Hill on Computer Engineering, 1996. p. 21.8 - 21.12.

XIE, X.L., BENI, G. A Validity Measure for Fuzzy Clustering. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 13, n. 8, August 1991.

ZAHID, N., ABOUELALA, O., LIMOURI, M., ESSAID, A. Fuzzy Clustering based on K-Neasrest-Neighbours rule. **Fuzzy Sets and Systems**, v. 120, p. 239-247, 2001.