ALGORITMOS PARA RECONHECIMENTO DE PADRÕES

ARMANDO ANTONIO MONTEIRO DE CASTRO
PEDRO PAULO LEITE DO PRADO
Departamento de Engenharia Elétrica
Universidade de Taubaté

RESUMO

O objetivo principal desse trabalho foi o desenvolvimento de algoritmos para sistemas de reconhecimento de padrões com ênfase em técnicas de agrupamento. Inicialmente são apresentados os conceitos básicos sobre reconhecimento de padrões e, a seguir, desenvolve-se uma visão sistêmica do problema, discutindo as fases e métodos de abordagem de projeto do reconhecimento de padrões. Finalmente, apresenta os algoritmos comumente usados para técnicas de busca e agrupamento, aplicando alguns exemplos para ilustrar seus passos e desenvolvimento. Os algoritmos analisados e implementados foram: "Similaridade Máxima", "MaxiMin-Distância", "K-Means" e "ISODATA". O objetivo proposto inicialmente foi alcançado, ou seja, foram desenvolvidos com êxito os algoritmos "Similaridade Máxima", "MaxiMin-Distância", "K-Means" e "ISODATA". Os exemplos desenvolveram-se de maneira a apresentar de forma didática a implementação dos algoritmos nas amostras de padrão determinadas para cada caso. A medida de similaridade adotada para comparação de padrões foi a distância euclidiana. Os algoritmos foram escritos em linguagem C++ e MatLab. Procurando tornar mais amigável o acompanhamento dos algoritmos implementados e a apresentação dos resultados, foi desenvolvida uma interface gráfica, utilisando-se o software MatLab.

PALAVRAS-CHAVE**:** reconhecimento de padrões, algoritmos, agrupamento, similaridade máxima, maximin-distância, k-means, isodata.

INTRODUÇÃO

O reconhecimento de padrões é uma habilidade extremamente desenvolvida nos seres humanos e em alguns animais. O ser humano é hábil em reconhecer rostos, vozes, caligrafias e, até mesmo, estados de humor de pessoas conhecidas. Alguns animais também têm essa característica desenvolvida, tais como os cães farejadores, que vasculham bagagens em terminais de desembarque em busca de drogas. O grau de refinamento do reconhecimento de padrões, por parte do ser humano, pode chegar a ponto de distinguir uma pintura de um grande mestre daquela feita por um exímio falsário ou, ainda mais, pode estabelecer uma tomada de decisão por parte de um operador em um dia de grande movimento em uma bolsa de valores.

Assim sendo, pode-se dizer que padrões são os meios pelos quais o mundo é interpretado e, a partir dessa interpretação, elaboram-se atitudes e decisões. Percebe-se, também, que nos exemplos citados, tal facilidade no reconhecimento de padrões está diretamente vinculada aos estímulos aos quais o indivíduo foi exposto anteriormente. Isso leva a supor que a estrutura selecionada pela evolução biológica para desempenhar bem a tarefa de reconhecimento de padrões incorpora alguma forma de aprendizado e evolui com a experiência. O grande desafio proposto neste inicio de século é o de desenvolver máquinas que tenham tais comportamentos. Tarefas de reconhecimento de voz, de caligrafias e de textos impressos usando caracteres de tipos distintos estão em fase avançada de desenvolvimento, mas muito ainda precisa ser feito para que seu desempenho se assemelhe ao do ser humano. Algumas áreas que poderiam ser destacadas na aplicação do reconhecimento de padrões são: identificação através de impressões digitais e análise da íris, diagnósticos médicos (Steiner, 1995), análise de imagens aeroespaciais, visão computacional (Perelmuter et al., 1995), diagnósticos pré e pós-natal e certos diagnósticos de câncer (Aguiar, 2000), reconhecimento de voz,(Ferreira,1994) investigação da qualidade do papel industrial (Steiner, 1995), processamento de imagens, análise de peças para manutenção preventiva, análise de caracteres manuscritos (Prado, 1975), análise de eletrocardiogramas, sinais de radar, (Mascarenhas,1987) reconhecimento e identificação de cromossomos(Todesco, 1995) dentre outras.

CONCEITOS BÁSICOS EM RECONHECIMENTO DE PADRÕES

O estudo do reconhecimento de padrões pode ser dividido em duas categorias básicas: (Tou e Gonzales, 1981).o estudo de seres humanos e organismos vivos com o objetivo de se estabelecer os modos pelos quais os mesmos desenvolvem e aprimoram suas capacidades de reconhecimento de padrões e o desenvolvimento de teorias e técnicas objetivando a construção de máquinas ou dispositivos capazes de apresentar características semelhantes às dos seres humanos em reconhecerem padrões. O presente trabalho desenvolve-se abordando a segunda categoria.

Padrão e Classe

Entende-se por padrão as propriedades que possibilitam o agrupamento de objetos semelhantes dentro de uma determinada classe ou categoria, mediante a interpretação de dados de entrada, que permitam a extração das características relevantes desses objetos, (Tou e Gonzáles, 1981). Entende-se por classe de um padrão um conjunto de atributos comuns aos

objetos de estudo. Assim, reconhecimento de padrões pode ser definido como sendo um procedimento em que se busca a identificação de certas estruturas nos dados de entrada em comparação a estruturas conhecidas e sua posterior classificação dentro de categorias, de modo que o grau de associação seja maior entre estruturas de mesma categoria e menor entre as categorias de estruturas diferentes. Os dados de entrada são medidos por sensores e selecionados segundo o conteúdo de informações relevantes para a decisão, e passam por um processo de redução de sua dimensionalidade para que possam ser usados pelo classificador, que o designará à classe que melhor o represente.

Fases do Reconhecimento de Padrões

Um sistema para reconhecimento de padrões engloba três grandes etapas: representação dos dados de entrada e sua mensuração, extração das características e finalmente identificação e classificação do objeto em estudo. A primeira etapa refere-se à representação dos dados de entrada que podem ser mensurados a partir do objeto a ser estudado. Essa mensuração deverá descrever padrões característicos do objeto, possibilitando a sua posterior classificação numa determinada classe. O vetor que caracteriza perfeitamente um objeto seria de dimensionalidade infinita, descrito por um vetor X:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ . \\ . \\ . \\ x_N \end{pmatrix}$$

onde: x_1, x_2, x_3,...,x_N são suas características. A segunda etapa consiste na extração de características intrínsecas e atributos do objeto e conseqüente redução da dimensionalidade do vetor padrão. É a fase da extração das características. A escolha das características é de fundamental importância para um bom desempenho do classificador. Esta escolha é feita objetivando os fenômenos que se pretende classificar. Exige-se, portanto, um conhecimento específico sobre o problema em estudo. Nesta etapa, os objetivos básicos são: a redução da dimensionalidade do vetor característico, sem que isso implique em perda de informação que possa ser relevante para a classificação, objetivando a redução do esforço computacional e a seleção das características significativas para a tarefa de classificação. A terceira etapa em reconhecimento de padrões envolve a determinação de procedimentos que possibilitem a identificação e classificação do objeto em uma classe de objetos. De modo diferente da segunda etapa, aqui a concepção do

classificador pode ser abordada de forma abstrata e independente da natureza do problema, pois, os métodos usados em reconhecimento de voz, análise de imagens, processamento de sinais de radar, inspeção de materiais, visão por computador ou detecção de avarias são muitas vezes os mesmos, possibilitando a aplicação dessas técnicas em contextos variados, sem perda de sua eficiência (Marques, 1999). Num sistema físico qualquer existe um número infinito de características que definem os padrões nele existentes. O Extrator de Características tem como função determinar e extrair as características mais significativas que contribuam para a descrição do objeto, dentre as infinitas características que possam descrevê-lo. Outro dado relevante é que o extrator de características varia com o sistema a ser analisado. A tabela 1 exemplifica várias tarefas de classificação, propostas por um sistema de reconhecimento de padrões, com seus dados de entrada e respectivos dados de saída:

Tabela 1: Exemplos de tarefas de classificação

Tarefas de Classificação	**Dados de Entrada**	**Dados de Saída**
Reconhecimento de caracteres	Sinais ópticos	Nome do caractere
Reconhecimento de Voz	Voz	Identificação da palavra
Diagnósticos Médicos	Sintomas	Identificação da Patologia
Previsão do tempo	Mapas atmosféricos	Chuva, Sol etc.

Uma vez extraídas as características é necessária a classificação do objeto. Esta classificação pressupõe a designação do objeto à uma determinada classe, dentre as várias que se apresentam. Nesta etapa o classificador "aprende" a distinguir dentre as classes, aquela à qual o objeto pertence. Padrões de uma mesma classe aglomeram-se em agrupamentos S_j. Se o treinamento do classificador exigir amplo conhecimento "a priori" da estrutura estatística dos padrões a serem analisados e o padrão de entrada for identificado como membro de uma

classe pré-definida pelos padrões de treinamento, o classificador será chamado de Classificador Paramétrico e a classificação se processa de forma supervisionada. Por outro lado, se o classificador utilizar determinado modelo estatístico, ajustando-se mediante processos adaptativos e a associação entre padrões se fizer com base em similaridades entre os padrões de treinamento, o classificador será chamado de Classificador Não-Paramétrico e a classificação se processará de forma não supervisionada.

A grande dificuldade na implementação de um projeto de reconhecimento de padrões está justamente na escolha da técnica adequada para que as fases do reconhecimento de padrões ocorram de modo a representar satisfatoriamente os fenômenos do mundo real.

A figura 1 ilustra, de forma mais detalhada, as diversas fases do reconhecimento de padrões.

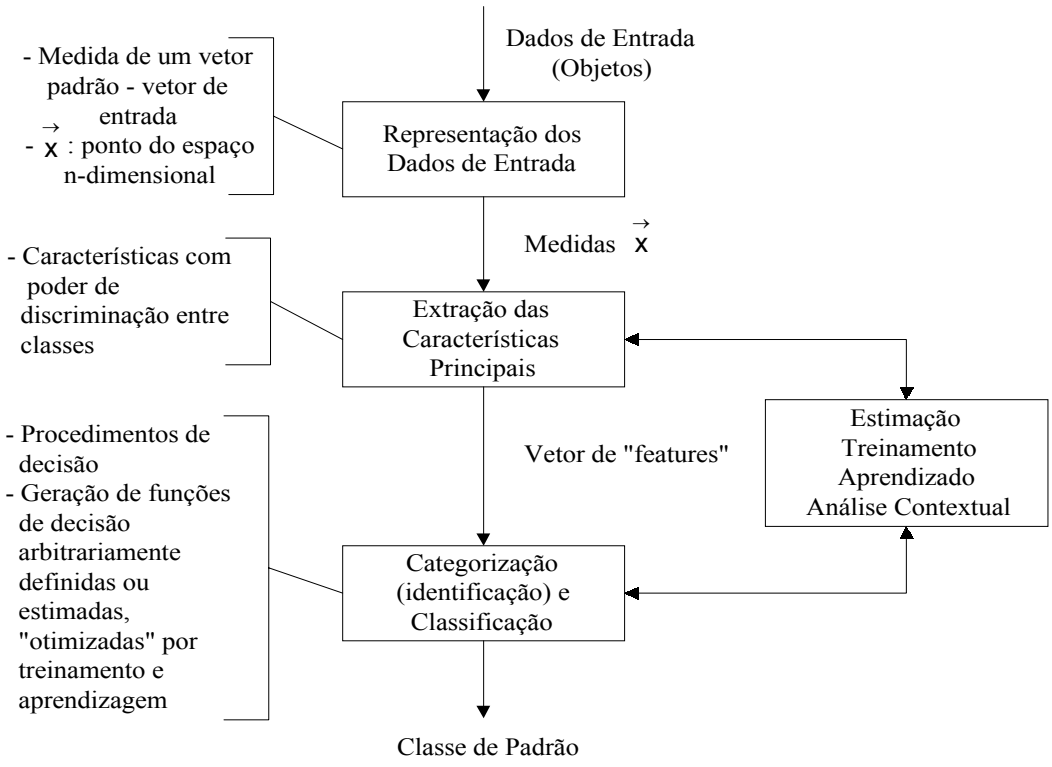

Figura 1: Fases do Reconhecimento de Padrões

Formas de Abordagens de Projeto de Reconhecimento de Padrões

As abordagens de projeto de reconhecimento de padrões são desenvolvidas em função da forma como as classes de padrão são categorizadas ou definidas. Basicamente são três as abordagens de projeto: rol de membros ("template matching"), propriedades comuns ("feature matching") e agrupamento ("clustering matching").

Rol de Membros ("Template Matching")

A caracterização de padrões mediante uma lista de padrões faz-se através de comparações com um modelo previamente armazenado cujas características servem de parâmetro para a comparação. Evidentemente, é um procedimento bastante elementar, podendo, em certos casos, acarretar classificações inadequadas em função de certos ruídos durante o processo de extração das características e posterior comparação com o padrão armazenado. Como exemplo de ruídos, em reconhecimento de caracteres, pode-se citar a impressão de um caracter mal delineado ou então uma maior ou menor porosidade do papel em que foi impresso, ocasionando concentração ou dispersão da tinta. É útil nas aplicações em que os padrões a serem comparados são bastante próximos do modelo padrão. A figura 2 ilustra a abordagem de projeto mediante a aplicação do conceito "Rol de Membros".

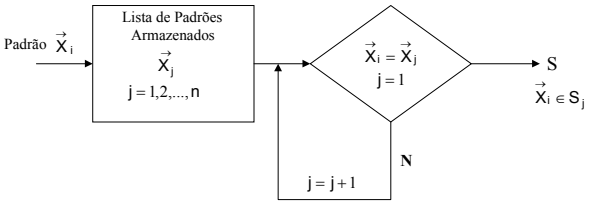

Figura 2: Abordagem do Rol de Membros ("template matching")

Propriedades Comuns ("Feature Matching")

A caracterização de padrões por propriedades comuns efetiva-se mediante algumas "características principais" inerentes aos elementos desta classe. Padrões pertencentes a uma mesma classe possuirão propriedades comuns de discriminação dessa classe. Desta forma, quando um padrão desconhecido é observado pelo sistema, suas características são extraídas e comparadas com aquelas armazenadas como discriminantes das classes. O classificador então, "classificará" este novo padrão em uma das classes existentes ou então designará o objeto a uma nova classe. O principal aspecto desta abordagem refere-se ao extrator das características do sistema, pois, dele dependerá o

bom desempenho do reconhecimento de padrões. Se todas as características de um padrão de classe são determinadas a partir de uma amostra, o processo de reconhecimento reduz-se simplesmente a estabelecer comparações com os novos objetos submetidos para análise. É extremamente difícil, porém, determinar completamente todas as características determinantes de uma classe de padrão. A utilização deste conceito implica freqüente desenvolvimento de técnicas que permitam aperfeiçoar e otimizar a extração das características dos objetos em estudo.

Agrupamento ("Clustering Matching")

Quando os padrões de uma classe são vetores cujos componentes são números reais, a classe do padrão pode ser estabelecida segundo formas do agrupamento, "clusters", desses pontos no plano. Havendo uma separação entre os pontos de forma clara, técnicas simples podem ser empregadas, tal como, "distância-mínima". Caso haja superposição entre os "clusters", técnicas mais elaboradas são necessárias, tais como, classificação por funções similares (métodos estatísticos), por treinamento de padrão (métodos determinísticos) ou outros algoritmos mais adequados.

Métodos de Reconhecimento de Padrões

As abordagens de projeto apresentadas anteriormente são implementadas através de três métodos de reconhecimento de padrões: métodos matemáticos, métodos lingüísticos ou sintáticos e métodos heurísticos. Não é raro o emprego de uma combinação dos métodos acima citados para que se efetive o reconhecimento de padrões (Tou e Gonzales, 1981).

Métodos Matemáticos

Os métodos matemáticos utilizam regras de classificação que são formuladas a partir de modelagem matemática dos conceitos de "propriedades principais" e "agrupamento". São subdivididos em duas categorias principais: métodos determinísticos e métodos estatísticos. O classificador de padrões estatístico baseia-se na "Regra de Classificação de Bayes", empregada quando a função densidade de probabilidade das classes de padrão e a probabilidade de ocorrência de cada classe de padrão são conhecidas. Assim, a principal tarefa para a utilização da classificação estatística é a estimação paramétrica das funções densidade de probabilidade ou, se necessário, aproximação funcional dessas densidades de probabilidade. Se forem conhecidas apenas a média e a variância das distribuições, a função gaussiana será utilizada. Os métodos

determinísticos têm formulação simplesmente matemática.

Métodos Lingüísticos ou Sintáticos

A abordagem de projeto é a de propriedades comuns, "feature matching". Baseia-se na caracterização de padrões por meio de uma estrutura hierárquica de elementos primitivos (sub-padrões) e suas relações. Desenvolve-se de forma semelhante à estrutura sintática de linguagens. É particularmente útil para padrões que não podem ser descritos convenientemente através de medidas numéricas, ou então, são tão complexos que suas características específicas não podem ser enunciadas, transformando-se então, em características globais.

Métodos Heurísticos

Baseiam-se em procedimentos "*ad hoc"* para tarefas especializadas de reconhecimento de padrões. Levam em consideração a intuição e experiência do programador em utilizar o conceito de rol de membros e propriedades comuns. Embora a abordagem heurística seja um importante método no reconhecimento de padrões, pouco pode ser dito sobre princípios generalizados nesta área, visto que cada problema requer a aplicação de regras especificas e elaboradas para tal fim. Por conseguinte, o desempenho de um sistema de reconhecimento de padrões baseado neste método dependerá de profundo conhecimento e intuição do elaborador do projeto.

Reconhecimento de Padrões Quanto à Supervisão

Reconhecimento de Padrões Supervisionado

Padrões representativos de cada classe estão disponíveis e o sistema é "ensinado" a reconhecer padrões por meio de esquemas de adaptação. Consiste, pois, na disponibilidade de "padrões de treinamento" e de um "procedimento de aprendizado". Exemplos de algoritmos utilizados no reconhecimento supervisionado são: *perceptron*, gradiente, erro quadrático mínimo, funções potenciais, etc .

Reconhecimento de Padrões Não-Supervisionado

São disponíveis apenas padrões de treinamento de classificação desconhecida. O reconhecimento de padrões de forma não-supervisionada é empregado quando não existe informação "a priori" acerca das classes dos protótipos. Os métodos para associar um dado agrupamento a cada padrão seguem algum critério de similaridade e são dependentes do algoritmo

empregado, dos dados utilizados e da medida de similaridade adotada.

ALGORITMOS DE BUSCA E AGRUPAMENTO

As técnicas de busca de agrupamento são empregadas quando os padrões das classes são números reais que se "agrupam" no espaço n-dimensional. São utilizados como ponto inicial para o reconhecimento de padrões não-supervisionado em cujo caso os padrões das classes são conhecidos a priori. Os centros dos agrupamentos obtidos por essa técnica podem ser interpretados como os diferentes padrões de classe, através das quais o treinamento pode ser realizado.

Classificação de Padrões Baseado em Similaridade Máxima.

A classificação por distância de funções é um dos primeiros conceitos em reconhecimento automático de padrões (Tou e Gonzáles, 1981). Esta técnica é uma ferramenta efetiva para a solução de problemas em que cada padrão de classe apresenta de modo claro, limitado grau de variabilidade, por exemplo, a identificação de caracteres impressos com tinta magnética em códigos de barra. Sob estas condições, similaridade máxima constitui uma abordagem adequada para o problema de classificação.

- Sejam M os agrupamentos de padrões de classes, representados pelos protótipos de padrão $Z_1, Z_2, Z_3 ... Z_M$.
- A distância Euclidiana entre padrões quaisquer x e cada protótipo Z_i será dada por:

$$D_i = \|x - Z_i\| = \sqrt{(x - Z_i)'(x - Z_i)} \qquad (3.2.1)$$

onde $(x - Z_i)'$ significa a transposta de $(x - Z_i)$. A classificação por similaridade máxima estabelece a distância entre um padrão x de classificação desconhecida em relação ao protótipo de cada classe e nomeia o padrão à classe que está mais próximo. Em outras palavras, x será designado para a classe w_i se:

$$D_i < D_j, \ \forall \ j \neq i$$

A equação (3.2.1) pode ser desenvolvida como:

$$D_i^2 = \|x - Z_i\|^2 = (x - Z_i)'(x - Z_i)$$

$$= x'.x - x'Z_i - Z'_i x + Z'_i Z_i$$

Sendo $x'Z_i = Z'_i x$, resulta:

$$D_i^2 = x'x - 2x'Z_i + Z'_i Z_i$$

$$= x'x - 2[x'Z_i - \frac{1}{2}Z'_i Z_i] \qquad (3.2.2)$$

Escolher o menor D_i^2 é equivalente a escolher o menor D_i, pois todas as distâncias são positivas. De (3.2.2) tem-se que o termo $x'x$ é independe de i para todo D_i^2, $i = 1,2,...,M$. Assim escolher o menor D_i^2, corresponde pois, escolher o máximo $[x'Z_i - \frac{1}{2}Z'_i Z_i]$. Define-se assim a função de decisão $d_i(x)$:

$$d_i(x) = x'Z_i - \frac{1}{2}Z_i'Z_i \text{ , } i = 1,2,...,M \qquad (3.2.3)$$

onde o padrão x é designado para a classe W_i se $d_i(x) > d_j(x) \quad \forall\ j \neq i$.

Sendo $d_i(x)$ uma função de decisão linear, isto é, se Z_{ij} , $j = 1,2,3,...,n$, são os componentes de Z_i e $W_{ij} = Z_{ij}$, onde $j = 1,2,3,...,n$.

$$w_{i,n+1} = -\frac{1}{2}Z_i'Z_i \qquad (3.2.4)$$

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \\ 1 \end{pmatrix}$$

podendo-se representar a equação (3.2.3) na forma linear

$$d_i(x) = w'_i x \text{ , com } i = 1,2,3,...,M \qquad (3.2.5)$$

onde: $W_i = (w_{i1}, w_{i2},..., w_{i,n+1})'$.

Classificação por Limiar Simples - "*Threshold Algorithm*"

O algoritmo Limiar Simples tem uma conceituação bastante simples.

Arbitrariamente é escolhida uma das amostras de padrão, disponíveis para análise, para ser o primeiro centróide de classe Z_1.

A seguir, são calculadas as distâncias entre o centróide Z_1 e cada e uma das amostras. Se essas distâncias forem menores que um limiar "T", previamente estabelecido, a amostra é designada à classe S_1 de centróide Z_1. Caso alguma distância seja maior que o limiar "T", novo centróide de classe é determinado e novamente são calculadas e comparadas as distâncias entre centróides e amostras com o limiar "T". Algumas características desse algoritmo merecem especial atenção. A escolha do centróide inicial, protótipo de classe, pode afetar de maneira significativa a classificação das amostras. Seu desempenho pode ser melhorado se os valores para o limiar "T" adotados forem próximos aos centróides de agrupamento. O valor arbitrado para o limiar "T" afeta a alocação dos padrões às classes de agrupamento. Caso esse valor seja muito grande, distintas amostras de padrão podem agrupar-se em uma mesma classe.

Por outro lado, se o valor adotado for muito pequeno as amostras constituirão inúmeras classes cujos padrões podem ter mesmas características. Em ambos os casos, tem-se a descaracterização da análise efetuada. A escolha do melhor valor para o limiar "T" dependerá de análise da disposição das amostras, da quantidade de amostras e de experimentações com vários valores (Feucht, 1977).

Algoritmo Maximin-Distância

O algoritmo Maximin-Distância é um procedimento heurístico baseado no conceito de medida de similaridade, por exemplo, a distância euclidiana. O algoritmo é similar, em princípio, ao apresentado no item 3.2, diferindo deste por identificar inicialmente as amostras mais afastadas entre si.

3.4 Algoritmo K-Means

O algoritmo k-means baseia-se na minimização de uma medida de custo, a distância interna entre os padrões de um agrupamento. A minimização do custo garante encontrar um mínimo local da função objetivo, que dependerá do ponto inicial do algoritmo. Esse tipo de algoritmo é chamado de "não-convexo", pois, a cada iteração diminui o valor da distorção, visto que o resultado final depende do ponto inicial usado pelo algoritmo. Uma derivação do algoritmo k-means, com heurísticas para divisão e união de "clusters", é o algoritmo ISODATA, que é o algoritmo k-means modificado. Apesar dessas características, o ISODATA também não é convexo e variante, ou seja, enfrenta problemas quando se depara com mínimos locais, durante a minimização do custo (Aguiar, 2000). Observe-se que: *" dado um conjunto S, distorção é qualquer aplicação d: S X S* $\rightarrow \Re_0^+$ *que associa cada par de elemento de S a um número real que traduz o afastamento entre esses elementos"* (Marques, 1999).

O Algoritmo

I. *Inicialização: Selecionar arbitrariamente os k centros iniciais de agrupamentos (centróides),* $Z_1, Z_2, ..., Z_k$.

II. *Determinar a distância* x_i, *com i = 1, 2, ...,k, entre cada centróide e os demais padrões. A designação dos padrões às classes de padrão se fará, no caso da distância euclidiana, quando as amostras de padrão atenderem à relação:*

$$\|x_i - Z_j(k)\| < \|x_i - Z_i(k)\| \quad , com$$

$$i = 1,2,...,k \text{ e } i \neq j .$$

A expressão $\|x - Zi\|$ *define a distância*

euclidiana entre os padrões, (amostras) e os

centróides. Formam-se k agrupamentos (cluster), compostos pelos elementos x_i *pertencentes à classe* $S_j(k) = \{x_1, x_2, ...\}$

III *Atualizar os centros de agrupamento a partir do passo 2, usando a relação:*

$$Z_j(k+1) = \frac{1}{N}\sum_{x \in S_j} x$$ *, onde N é numero de elementos de* S_j*.. Esta atualização minimiza a soma do quadrado das distâncias de cada padrão do agrupamento ao novo centro de agrupamento.*

IV *Comparar os centróides* $Z_j(k+1)$ *com o centróide* $Z_j(k)$. *Se* $Z_j(k+1) \neq Z_j(k)$ *então novos centróides ficam determinados e*

repete-se o passo 2 com $Z_j(k+1)$ *no lugar de* $Z_j(k)$*, definindo assim os novos elementos de cada classe. Se* $Z_j(k+1) = Z_j(k)$ *o algoritmo converge e o procedimento está terminado, com a designação de cada padrão á sua respectiva classe de padrões.*

O comportamento do algoritmo k-means apresenta vantagens no que concerne a simplicidade e eficiência. É rápido para cálculos simples, possibilitando o processamento seqüencial dos dados, acarretando baixa quantidade de armazenamento de informações a serem processadas. A desvantagem é a sua dependência dos valores iniciais de k, da ordem em que as amostras são processadas, da escolha dos primeiros centros de agrupamento e da geometria das amostras disponíveis para análise. Em alguns casos sua utilização requer experimentação com vários valores de k e diferentes escolhas dos parâmetros iniciais(Duda, 1974).

Algoritmo ISODATA

O algoritmo ISODATA, *Iterative Self-Organizing Data Analysis Techniques,* é um algoritmo semelhante, em princípio, ao algoritmo k-means. Desenvolve-se baseado em critérios de minimização e aplicação de métodos heurísticos em um procedimento iterativo para determinação de centróides de classes de padrões. Seu desenvolvimento segue os seguintes passos:

1. *Inicialização: Selecionar de forma arbitrária os centros de agrupamento* $Z_1, Z_2, ..., Z_{Nc}$. N_c*. O número de centros de agrupamentos arbitrado,*

necessariamente, não precisa ser igual ao número final de centróides obtidos após o término do algoritmo.

2. *Estabelecer os valores dos parâmetros iniciais:*

   K = Número de agrupamentos (clusters) desejado

   θ_N = *Parâmetro mínimo de amostras permitido por classe, usado na eliminação de agrupamentos*

   θ_S = *Parâmetro a ser comparado com o desvio padrão, usado na divisão de agrupamento*

   θ_C = *Parâmetro de agrupamento, usado para agregar agrupamentos*

   L = Número máximo de amostras que podem juntar-se a uma classe durante o desenvolvimento do algoritmo.

   I = Número máximo de iterações permitidas.

3. *Designação das amostras de padrões aos centros de agrupamento mais próximo, usando a relação:*

   $x \in S_j$ se $\|x - Z_j\| < \|x - Z_i\|$ *com* $i = 1,2,...,N_c$, $i \neq j$

4. *Descartar os conjuntos de padrões S_j que não apresentarem quantidade mínima de amostras ou seja, para as classes S_j em que $N_j < \theta_N$, descartar S_j e reduzir o número de classes em uma unidade, $N_C = N_C - 1$.*

5. *Atualizar os centros de agrupamento Z_j usando a relação:*

   $$Z_j = \frac{1}{N_j} \sum_{x \in S_j} x \, , \; j = 1,2,,...,N_c$$ *onde N_j é o número de amostras presentes nas respectivas classes.*

6. *Calcular a dispersão $\bar{D}_j$, distância média de cada amostra x, da classe S_j , ao centro de agrupamento Z_j correspondente, usando a relação:*

   $$\bar{D}_j = \frac{1}{N_j} \sum_{x \in S_j} \|x - Z_j\| \; , \; j = 1,2,...,N_c$$

7. *Calcular a dispersão global $\bar{D}$, distância média global entre amostras e respectivo centróide, usando:*

   $$\bar{D} = \frac{1}{N} \sum_{j=1}^{N_c} N_j \bar{D}_j$$

   onde: N é a quantidade de amostras disponíveis para análise e N_j é a quantidade de amostras presentes na classe S_j. $\bar{D}$ fornece o "espalhamento" entre mostras de padrão de uma mesma classe, usado como parâmetro na

divisão de classes.

8. *Terminar, dividir ou reagrupar:*
   a) *se for a última iteração, fazer* $\theta_C = 0$ *e ir para o passo 12.*
   b) *Se* $N_c \leq \frac{k}{2}$ *, ir para o passo 9 (quantidade de centróides menor que o esperado, buscar divisão de classes).*
   c) *Se for uma iteração par ou* $N_c \geq 2k$ *ir para o passo 12 (quantidade de centróides acima do esperado, buscar agrupamento de classes), caso contrário continue*

9. *Calcular o vetor desvio padrão* σ_j *para cada amostra de padrão em relação aos eixos coordenados,* $\sigma_j = (\sigma_{1j}, \sigma_{2j}, ..., \sigma_{nj})'$, *usando a relação*

$$\sigma_{ij} = \sqrt{\frac{1}{N_j} \sum_{x \in Sj} \left(x_{ik} - Z_{ij}\right)^2}$$

$i = 1,2,...n \; ; j = 1,2,..., N_c ; k = 1,2,..., N_j$

onde:

n = Dimensionalidade da amostra

x_{ik} = *i-ésimo componente da k-ésima amostra em* S_j

Z_{ij} = *i-ésimo componente de* Z_j

N_j = *número de amostras da classe* S_j

Cada componente de σ_j *representa o desvio padrão das amostras em* S_j *ao longo do principal eixo coordenado.*

10. *Calcular o componente máximo da dispersão* σ_j *e denotá-lo* $\sigma_{j_{máx}}$, $j = 1,2,...,N_c$,

11. *Divisão: Se para algum* $\sigma_{j_{máx}}$, $j = 1,2,...,N_c$, *tem-se* $\sigma_{j_{máx}} > \theta_S$ *e:*
   a) $\overline{D}_j > \overline{D}$ *e* $N_j > 2(\theta_N + 1)$ *ou*
   b) $N_c \leq \frac{k}{2}$

Dividir a classe S_j, *de centro em* Z_j, *em duas outras classes cujos centros serão dados por* Z_j^+ e Z_j^- *, obtidos em função de* γ_j, *com* $\gamma_j = \kappa\sigma_{jmáx}$ *e* $0 < \kappa \leq 1$, *eliminando-se* Z_j *e incrementando* N_C *de 1 unidade.*

Assim, tem-se: $Z_j^+ = f(Z_j, \gamma_j)$ *e* $Z_j^- = f(Z_j, \gamma_j)$.

Ocorrendo a divisão vá para o passo 3, caso contrário continue.

12. Calcular os pares de distâncias D_{ij} *entre os centros de agrupamentos,* $D_{ij} = \|Z_i - Z_j\|$ *;* $i = 1,2,...,N_c - 1$ *e* $j = i+1,...,N_c$ *, e comparar* D_{ij} *com* θ_N*. Reordenar as L menores distâncias que* θ_N*,* $(\overline{D}_{ij} < \theta_c)$ *, em ordem crescente:* $\lfloor D_{i1j1}, D_{i2j2},...,D_{iLjL} \rfloor$ *onde* $D_{i1j1} < D_{i2j2} < .,D_{iLjL}$ *e L é o número máximo de classes que podem se agrupar.*

13. União dos Agrupamentos: para cada distância D_{iljl} *é associado um par de agrupamentos de centros* Z_{il} *e* Z_{jl}*. Iniciar o agrupamento com a menor dessas distâncias. Para l = 1,2,...,L , unir as duas classes usando a relação:*

$$Z_l^* = \frac{1}{N_{il} + N_{jl}} \left[N_{il}(Z_{il}) + N_{jl}(Z_{jl}) \right],$$

descartando Z_{il} *e* Z_{jl} *e reduzindo* N_C *de uma unidade.*

14. Terminar ou repetir:

a) Se for a última iteração, o algoritmo segue para o passo 15.

b) Se não for última iteração, pode ser necessária a troca de alguns dos parâmetros iniciais, o algoritmo segue para o passo 2.

c) Se não for última iteração e não houver necessidade de alteração nos parâmetros inicias. O algoritmo segue para o passo 3. Observe-se que uma iteração é considerada cada vez que o algoritmo retorna ao passo 2 ou 3.

14. O algoritmo está encerrado, com a designação dos padrões às respectivas classes de padrão.

RESULTADOS E CONCLUSÕES

O objetivo principal proposto inicialmente foi alcançado, ou seja, foram desenvolvidos com êxito os algoritmos para sistemas de reconhecimento de padrões, com ênfase em técnicas de agrupamento, "Similaridade Máxima", "MaxiMin-Distância", "K-Means" e "ISODATA".

Os exemplos desenvolveram-se de maneira a apresentar de forma didática a implementação dos algoritmos nas amostras de padrão determinadas para cada caso.

Foram escolhidas amostras com vetor de características bidimensional dada a maior facilidade de visualização dos resultados obtidos

após a implementação dos algoritmos. A medida de similaridade adotada para comparação de padrões foi a distância euclidiana dada por

$$D_i = \|x - Z_i\| = \sqrt{(x - Z_i)'(x - Z_i)}$$

onde $(x - Z_i)'$ significa a transposta de $(x - Z_i)$. Os algoritmos foram escritos em linguagem C++(Borland C++ for Windows, V5.02). Objetivando uma forma mais amigável de apresentação dos resultados dos algoritmos, foi desenvolvida uma interface gráfica, (Castro, 2001) usando-se o software Matlab, em sua versão 5.3, para o algoritmo K-Means. Os exemplos apresentados foram limitados ao espaço bidimensional, objetivando ilustrar de forma didática os fundamentos teóricos desenvolvidos e tornar possível seu desenvolvimento passo a passo, embora, usualmente, as tarefas de reconhecimento de padrões contemplem variáveis n-dimensionais, $n \geq 3$.

ABSTRACT

This work aims at developing algorithms for pattern recognition, mainly on clustering techniques. Initially, the basic concepts of pattern recognition are presented. Then, a systemic vision of the problem is developed, with the discussion of the approaches for the project of pattern recognition. Finally, the following algorithms for cluster-seeking are implemented: "Maximim Similarity", "Maximin-Distance", "K-Means" and "ISODATA". An user-friendly interface helps to present the programs, written in C++ and in Matlab.

KEY WORDS*:* pattern recognition, algorithms, clustering, maximim similarity, maximin-distance, k-means, isodata.

REFERÊNCIAS BIBLIOGRÁFICAS

Aguiar, R. G. *Segmentação de Imagens em Cores de Himunohistoquímica.* Dissertação de Mestrado. Universidade Federal de São Carlos, 2000.

Castro,A.A.M. *Algoritmos Para Reconhecimento de Padrões*. Dissertação de Mestrado, Universidade de Taubaté, 2001.

Duda, R. O., Hart, P. E. *Pattern Classification and Scene Analysis*. Wiley -Interscience, New York. 1974.

Ferreira, M. F**.** *O Reconhecimento de Padrões.* Dissertação de Mestrado. Universidade de Brasília, 1994.

Feucht, D. *Pattern Recognition, Basic Concepts and Implementations*. Computer Design, 1977**.**

Marques, J,S. *Reconhecimento de Padrões Métodos Estatísticos e Neuronais.* IST Press, Portugal 1999.

Mascarenhas, N. D. A**.** *Breve Introdução ao Reconhecimento Estatístico de Padrões.* 39ª Reunião Anual da SBPC. 1987.

Perelmuter, G., Carrera, E. V., Vellasco, M., Pacheco, A. *Reconhecimento de Imagens Bidimensionais Utilizando Redes Neurais Artificiais.* Anais do VII SIBGRAPI, p. 197-203, 1995.

Prado, A. Jr., Elfes, A. *Um Projeto em Reconhecimento de Padrões de Forma.* Monografia de Graduação, ITA, São José dos Campos, SP, 1975.

Steiner, M. T. A. *Uma Metodologia para o Reconhecimento de Padrões Multivariados com Resposta Dicotômica.* Tese de Doutorado. Florianópolis-SC, 1995.

Todesco, J. L. *Reconhecimento de Padrões usando Rede Neuronal Artificial com uma Função de Base Radial: uma aplicação na classificação de cromossomos humanos.* Tese de Doutorado. Florianópolis-SC, 1995.

Tou, J. T., Gonzalez, R. C. *Pattern Recognition Principles.* Addison-Wesley Publishing Company, Massachusetts, 1981.