COMPARAÇÃO ENTRE OS MÉTODOS DE ENTROPIA E DA DISTÂNCIA DE JEFFREYS-MATUSITA EM PROBLEMAS DE SELEÇÃO DE ATRIBUTOS

F. A. Mitsuo Ii, L. V. Dutra e C. L. Mendes

Instituto de Pesquisas Espaciais

Conselho Nacional de Desenvolvimento Científico e Tecnológico

Caixa Postal 515, 12200 - São José dos Campos, SP, Brasil

RESUMO

A pesquisa teve por objetivo avaliar o desempenho do método da Distância J-M e da Entropia, como critérios de seleção de atributos, utilizando-se imagens do satélite LANDSAT. Selecionou-se como área de estudo uma imagem na região de Ribeirão Preto, São Paulo, com predominância de cana. A partir dos 4 canais originais do satélite LANDSAT, extrairam-se mais 8 canais, utilizando-se filtragens passa-baixa e passa-alta para gerar atributos espaciais. Definiram-se 5 classes de treinamento para aquisição dos parâmetros necessários. Dos 12 canais obtidos, escolheram-se 4, segundo o critério da Distância J-M e da Entropia, número esse definido pela capacidade de manipulação do imageador I-100 e pelo custo computacional. Fez-se a avaliação, obtendo-se as matrizes de classificação para as áreas-teste e as de treinamento com a utilização do classificador de máxima verossimilhança com hipótese gaussiana. A partir das matrizes de classificação, extraíram-se índices de desempenho que mediram a precisão do classificador, para aquele conjunto de classes fixado e para cada conjunto de atributos, selecionados segundo cada critério. Os resultados mostraram que, com atributos espaciais e classificação supervisionada, o critério da Entropia é melhor, pois permite uma definição mais precisa e generalizada das classes. Todavia, o critério da Distância J-M reduz fortemente o erro de classificação nas áreas de treinamento.

ABSTRACT

This research had the purpose of evaluating the performance of entropy and JM distance feature selection methods, using LANDSAT satellite images. A study area near Ribeirão Preto in São Paulo state was selected, with predominance in sugar cane. Eight features were extracted from the 4 original bands of LANDSAT image, using low-pass and high-pass filtering to obtain spatial features. There were 5 training sites in order to acquire the necessary parameters. Two groups of four channels were selected from 12 channels using JM distance and Entropy criterion. The number of selected channels was defined by physical restrictions of the image analyzer and computational costs. The evaluation was perfomed by extracting the confusion matrix for training and tests areas, with a maximum likelihood classifier, and by defining performance indexes based on those matrixes for each group of channels. The results showed that with spatial features and supervised classification, the entropy criterion is better in the sense that allows a more accurate and generalized definition of class signature. On the other hand, JM-distance criterion strongly reduces the misclassification within training areas.

1. INTRODUÇÃO

Desde o advento dos computadores digitais, tem havido um constante esforço no sentido de idealizar métodos automáticos, que substituam o homem no trabalho de tomar decisões, muitas vezes monótono e repetitivo, ou que façam essa tarefa de maneira rápida e precisa.

Estudos intensivos de problemas de classificação - ato de associar um objeto físico ou evento a uma das várias categorias especificadas - têm conduzido à formulação de muitos modelos matemáticos que determinam a base teórica para o projeto de classificadores.

Como exemplo de problemas de classificação, podem-se citar: previsão numérica de tempo, diagnóstico de pacientes através da análise de eletrocardiogramas e raios X, reconhecimento de assinaturas escritas à mão, impressão digitais, etc.

Um sistema de classificação de padrões pode ser dividido em duas partes: o extrator de características e o classificador (Figura 1).

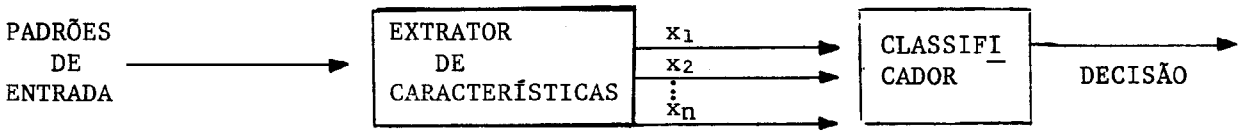

Fig. 1 - Sistema de classificação de padrões

O extrator de atributos tem a função de reduzir os dados naturais, medindo um certo conjunto de "atributos" ou "propriedades" que melhor caracterizem os objetivos de interesse. Esses atributos, ou mais precisamente os valores desses atributos, passam por um classificador que avalia as evidências apresentadas, segundo determinado critério, e associa uma categoria ao objeto.

O critério de classificação é, usualmente, a minimização do erro de classificação (ou erro de reconhecimento).

Recentemente, muitas técnicas de classificação têm sido propostas. Se as medidas características, que descrevem todos os possíveis padrões de entrada em cada classe, puderem ser caracterizadas por quantidades (funções) determinísticas ou estatísticas (isto é, distribuição de probabilidade), estas podem ser classificadas em técnicas de classificação estatísticas ou determinísticas.

De outra forma, se as propriedades das medidas características, que descrevem todos os padrões em cada classe, puderem ou não ser expressas em forma paramétrica (por exemplo, por uma função densidade de probabilidade de forma conhecida), estas podem ser classificadas em técnicas de classificação paramétricas ou não-paramétricas.

A seleção de uma técnica particular em aplicações práticas depende, às vezes, da natureza do problema, de uma informação disponível a priori, e da preferência do analista.

Supondo-se que existam M classes-padrão possíveis W_1, W_2,..., W_m, e N características x_1, x_2,..., x_n a serem extraídas para classificação, cada conjunto de N medidas características pode ser representado como um vetor N-dimensional $\vec{X} = [x_1, x_2, \ldots, x_n]$, ou como um ponto no espaço N-dimensional, chamado espaço característico Ω_x.

Usualmente, o uso de um grande número de medidas características aumentará a complexidade e o tempo computacional do classificador. Técnicas de seleção de atributos permitem selecionar um número menor de atributos, aumentando assim a eficiência das tarefas computacionais, sem prejudicar demasiadamente a precisão.

2. EXTRAÇÃO DE ATRIBUTOS

Um atributo de imagem é uma propriedade que pode ser medida. Os atributos naturais são aqueles que derivam da aparência da imagem, como o nível de cinza, bordas e textura. Os artificiais são aqueles obtidos por manipulações e por medidas na imagem, como histograma e frequência espacial.

O atributo mais utilizado em problemas de classificação de padrões são os níveis de cinza de imagens multiespectrais ou cena, comumente obtida a partir de satélites de recursos terrestres.

Mas cada componente de uma cena ou imagem não carrega apenas, como informação, o nível de cinza do ponto. O relacionamento do ponto com seus vizinhos mais próximos também é um tipo de informação denominada informação espacial.

Uma maneira de extrair atributos espaciais é a partir de filtragem linear ou não-linear (Dutra, 1982).

O atributo denominado variação (Schackter et alii, 1979), que é um tipo de filtro não-linear é aplicado a cada componente de uma cena e, para ser definido, considere-se a seguinte disposição dos pontos em uma vizinhança:

a	b	c
d	x	e
g	h	i

Daí obtém-se:

- variação horizontal:

$$HTV = |a-b| + |d-x| + |g-h| + |b-c| + |x-e| + |h-i| \quad (1)$$

- variação vertical:

$$VTV = |a-d| + |b-x| + |c-e| + |d-g| + |x-h| + |e-i| \quad (2)$$

- variação total:

$$TV = HTV + VTV \quad (3)$$

Esse filtro informa sobre a rugosidade local da imagem. As regiões com alta TV são regiões muito rugosas.

O filtro linear passa-baixa realiza uma média sobre a imagem e é utilizado normalmente para diminuir a influência do ruído no resultado do processo de classificação de padrões (Figura 2).

Por exemplo:

$$h_1 = \frac{1}{21}\begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Fig. 2 - Filtro passa-baixa (5x5) octogonal.

Esse filtro associa a cada ponto a média em uma região 5x5 octogonal ao redor dele.

Além dos atributos gerados a partir dos componentes de uma dada cena, pode-se utilizar, conjugada a ela, uma cena referente ao mesmo local, mas de outra data, ou então uma obtida por outro tipo de sensor ou plataforma.

Com isso gera-se, para representar uma dada região ou objeto, um conjunto com um grande número de atributos, sendo que alguns deles são referentes ao caráter espectral de resposta à radiação solar, outros ao caráter temporal de variação dessas respostas, ou ao caráter espacial de relacionamento entre os pontos da imagem.

Faz-se necessário então escolher, dentre os atributos gerados ou coletados, um subconjunto que satisfaça um requisito, normalmente minimização do erro de classificação.

3. MÉTODOS DE SELEÇÃO DE ATRIBUTOS

Existe um compromisso muito importante entre o número de atributos (canais) utilizados na classificação de um padrão e o tempo computacional.

A precisão de uma classificação será tanto maior quanto maior for o número de canais disponíveis utilizados. Entretanto, isso acarretará um número maior de operações e, consequentemente, mais tempo será exigido.

Desse compromisso surge o problema básico de seleção de atributos em classificação de padrões:

- Dado um conjunto de N canais, achar o melhor subconjunto de K canais a serem usados para classificação, os quais provêem um compromisso ótimo entre a precisão na classificação e o tempo/custo computacional.

O ideal seria resolver este problema, computando-se a probabilidade do erro de classificação associado a cada subconjunto de $\overline{K}$ canais e, então, selecionando-se aquele que produz o menor erro. Contudo, geralmente não é fácil realizar as operações exigidas, pois a integração numérica necessária para computar os erros é impraticável.

Como exemplo, considere-se que:

$$\begin{bmatrix} N \\ n \end{bmatrix} \triangleq \frac{N!}{n!(N-n)!}$$

subconjuntos de atributos devem ser avaliados. Assim, por exemplo, para selecionar os quatro melhores atributos entre os doze disponíveis, exigem-se:

$$\begin{bmatrix} 12 \\ 4 \end{bmatrix} = \frac{12!}{4!8!} = 495$$

integrações no espaço 4-dimensional. Mesmo em computadores muito rápidos, tais computações seriam proibitivas. Assim, métodos alternativos devem ser encontrados para seleção de atributos.

Uma aproximação que tem sido muito investigada baseia-se no conceito de uma medida de "distância estatística" entre as densidades de probabilidade, que caracterizam as classes padrão. (Ii, 1982).

O ideal seria obter uma medida de distância com a seguinte propriedade:

- Se a distância entre duas classes for maior para um conjunto de canais α do que para um conjunto de canais β, então a probabilidade de erro obtida para o conjunto α seria menor do que para o conjunto β.

Infelizmente, nenhuma das medidas de distância, que tem sido propostas, possui exatamente essa propriedade.

Contudo, diversas distâncias possuem a característica de terem limiares superior e/ou inferior para a probabilidade de erro a elas associados. Assim, se a distância entre duas classes for maior para um conjunto α de atributos do que para um conjunto β, então, o limiar inferior e/ou superior para a probabilidade de erro obtida para o conjunto α é menor do que para o conjunto β.

Pode-se observar que essa propriedade é subótima, pois não se está minimizando, diretamente, a probabilidade de erro associada, e sim, os limiares inferior e/ou superior para a probabilidade de erro.

Como exemplo de medidas de distância estatística, que possuem essa característica, pode-se citar a Divergência, a Divergência Transformada, a Distância de Bhattacharyya (Distância B) e sua relacionada Distância Jeffreys-Matusita (Distância J-M).

Swain e King (1973) realizaram diversos experimentos sobre os métodos de medida de distância estatística e concluíram que o critério da Distância J-M leva algumas vantagens sobre os outros métodos, com relação à previsão correta dos melhores atributos para o reconhecimento multiclasses.

A Distância B é função escalar das funções densidade de probabilidade de 2 classes e definida como:

$$B = - \ln \rho \text{ ou } \rho = e^{-B}, \qquad (4)$$

onde

ρ = coeficiente de Bhattacharyya, dada por:

$$\rho = \int_{-\infty}^{\infty} \left(p(\vec{X}/w_1)\ p(\vec{X}/w_2)\right)^{1/2} dx \qquad (5)$$

A Distância JM é dada por:

$$d_{JM}^2 = 2(1-\rho) \Longrightarrow d_{JM} = \left(2(1-\rho)\right)^{1/2} \qquad (6)$$

Para o caso de 2 classes, podem ser obtidos limites superiores e inferiores para a probabilidade de erro em função de ρ. Sendo P_E a probabilidade de erro, e P_1 e P_2 probabilidades a priori de w_1 e w_2, respectivamente, tem-se:

$$\frac{1}{4}\rho^2 \leq P_1P_2\rho^2 \leq \frac{1}{2}\left(1 - \sqrt{1 - 4P_1P_2\rho^2}\right) \leq$$

$$P_E \leq \sqrt{P_1P_2}\ \rho \leq \frac{1}{2}\rho \qquad (7)$$

Para densidades gaussianas a Distância B é dada por:

$$B = \frac{1}{8}(\vec{\mu}_1 - \vec{\mu}_2)^T \left(\frac{\sum_1 + \sum_2}{2}\right)(\vec{\mu}_1 - \vec{\mu}_2) + \frac{1}{2}\left\{\frac{\frac{1}{2}\left|\sum_1 + \sum_2\right|}{\left|\sum_1\right|^{1/2}\left|\sum_2\right|^{1/2}}\right\}, \qquad (8)$$

onde

$\vec{\mu}_1$ e $\vec{\mu}_2$ são vetores média e $\sum_i$ são matrizes de covariância para as classes 1 e 2.

É difícil derivar uma expressão semelhante para outros tipos de função densidade de probabilidade; sabe-se, no entanto que, para a maior parte dos casos de imagens naturais, o modelo gaussiano se ajusta satisfatoriamente.

Quando se tem 2 classes, basta escolher o subconjunto com M atributos, para o qual a distância JM é maior. Para o caso de mais de 2 classes, costuma-se aplicar dois critérios para a escolha do melhor subconjunto: um subconjunto é escolhido, para o qual a distância média entre as distâncias JM para todos os pares de classes é maximizada. Outro critério é utilizado considerando-se o subconjunto que tenha a maior das distâncias JM mínima entre os pares de classes.

Um método alternativo para seleção de atributos é o critério da Entropia.

Entropia é comumente interpretada como a incerteza média da fonte de informação. A quantidade média de informação obtida, ao se realizar uma observação numa fonte, é igual à incerteza média que se tinha antes dessa observação (Young and Calvert, 1974).

Para padrões gaussianos, a entropia é definida como:

$$H(x) = \frac{1}{2}\ln\left|\sum_i\right| + \frac{N}{2}\ln 2\pi e \qquad (9)$$

onde

$\left|\sum_i\right|$ = determinante da matriz de covariância da classe i

N = número de atributos.

No caso de multiclasses e considerando-se que as classes sejam independentes entre si, para escolher o subconjunto de atributos que maximize o conteúdo de informação, basta utilizar o conjunto que maximize o seguinte parâmetro:

$$S = \sum_{i=1}^{m} \ln\left|\sum_i\right|, \qquad (10)$$

onde se despreza os termos constantes, pois eles não influenciam a regra de decisão, e M é igual ao número de classes.

4. RESULTADOS

Os experimentos foram efetuadas utilizando-se uma imagem LANDSAT-C, órbita 78, ponto 27, de abril de 1978, sobre a área de Ribeirão Preto. Obtiveram-se também imagens tira-

das de avião sobre a mesma área, o que permitiu escolher as áreas testes e de treinamento com boa precisão (Dutra, 1982).

As classes utilizadas e o número de pontos nas áreas teste e de treinamento são apresentados na Tabela 1.

TABELA 1

CLASSES USADAS

	CLASSE	NÚMERO DE PONTOS	
		ÁREAS DE TREINAMENTO	ÁREAS TESTE
1	Cana	252	108
2	Cana nova	216	108
3	Pasto	108	72
4	Água	72	36
5	Infra-estrutura	72	36
6	Mata	72	36

No experimento foram utilizados 12 atributos, assim distribuídos:

- os atributos 1 a 4 são os canais originais 4 a 7 do LANDSAT.
- os atributos 5 a 8 são as médias em regiões 5x5 dos canais 4 a 7 do LANDSAT (convolução dos canais 4 a 7 com a máscara da Figura 2).
- os atributos 9 a 12 informam sobre a rugosidade local dos canais 4 a 7 do LANDSAT e são obtidos pela aplicação do operador variação (Equação 3). Esses canais são amaciados pela convolução com o filtro da Figura 2 para diminuir a influência do ruído.

A fim de comprovar a eficiência do método, foram obtidas as matrizes de classificação das áreas teste e de treinamento para os canais originais e os escolhidos pelo critério da Distância J-M e da Entropia.

O classificador utilizado é o de máxima verossimilhança (Velasco et alii, 1978), que é um classificador do tipo estatístico supervisionado (usa áreas de treinamento para aquisição dos parâmetros necessários).

As matrizes de classificação apresentam, de forma sucinta, o resultado da classificação de áreas de classificação conhecida a priori.

Os erros cometidos ao classificar incorretamente pontos de identidade conhecida permitem estimar os erros envolvidos.

A partir dessas matrizes foi possível obter o "desempenho médio" (Dm), definido como a média da percentagem de classificação correta de cada área teste ou de treinamento, ponderada pelo número de pontos de cada uma. A "abstenção média" (Am) foi definida como sendo a percentagem média de abstenção das áreas, ponderada pelo número de pontos delas. A "confusão média" (Cm) foi definida como sendo o erro médio, ponderado pelo número de pontos das áreas.

A Tabela 2 mostra a matriz de classificação para as áreas de treinamento, utilizando-se os canais originais para o limiar de classificação igual a 5 (L = 5).

Os canais selecionados pelo critério da máxima distância J-M média e da máxima distância J-M mínima coincidiram e são apresentados na Tabela 3.

Os canais selecionados pela máxima soma das entropias são os apresentados na Tabela 4.

Os resultados de Dm, Am e Cm para áreas teste e de treinamento, utilizando-se os canais escolhidos pelos 2 critérios estão consolidados nas Tabelas 5 e 6. Os resultados para os canais originais são apresentados a título de comparação.

Pode-se observar que os resultados das matrizes de classificação para áreas de treinamento são melhores para os canais selecionados por distância J-M com diminuição acentuada da confusão média, embora haja diminuição da Cm também para os canais escolhidos por entropia.

Para as áreas testes, o resultado foi muito melhor para os canais escolhidos por entropia; isso demonstra que a escolha por distância estatística é otimizada em relação às áreas de treinamento e, portanto, dependente muito delas, e os canais escolhidos por entropia apresentam maior grau de generalização ou extensão de assinatura.

5. CONCLUSÕES E SUGESTÕES PARA FUTURAS PESQUISAS

Em processos de extração de atributos por rotação espectral, prova-se que a transformação dos componentes principais transfere o máximo de informação (entropia) para os primeiros canais. Essa transformação tende também a manter a representação dos dados, pois ela minimiza o erro quadrático médio de representação.

Dos canais escolhidos para o critério de entropia, observa-se que eles são os que apre-

sentam maiores detalhes visuais, carregam a informação espacial e apresentam grande variância.

Dentre os canais escolhidos pelo critério da distância J-M estão os canais média, e como supõe-se que pontos contíguos pertençam a mesma classe, quando se usa a filtragem passa-baixa, os pontos aproximam-se da média tanto espacial quanto espectralmente, aumentando assim a distância estatística e diminuindo a probabilidade de erro de classificação.

TABELA 2

MATRIZ DE CLASSIFICAÇÃO PARA ÁREAS DE TREINAMENTO, UTILIZANDO-SE CANAIS ORIGINAIS COM L = 5

MATRIZ DE CLASSIFICAÇÃO							
	N*	1	2	3	4	5	6
1. Cana	0,8	97,6	0,0	1,6	0,0	0,0	0,0
2. Cana nova	0,5	0,0	92,1	0,0	0,0	7,4	0,0
3. Pasto	0,0	0,0	0,0	100,0	0,0	0,0	0,0
4. Água	0,0	0,0	0,0	4,2	95,8	0,0	0,0
5. Infra-estrutura	0,0	0,0	8,3	0,0	9,7	81,9	0,0
6. Mata	0,0	0,0	0,0	0,0	0,0	0,0	100,0
Desempenho Médio DM = 95,1 %							
Abstenção Média AM = 0,4 %							
Confusão Média CM = 4,5 %							

N* = Não-classificado

TABELA 3

SELEÇÃO 1

NÚMERO DOS CANAIS	DENOMINAÇÃO DOS CANAIS
5	Média (5x5) do canal 4 do LANDSAT
8	Média (5x5) do canal 7 do LANDSAT
9	Variação suavizada do canal 4 do LANDSAT
10	Variação suavizada do canal 5 do LANDSAT

TABELA 4

SELEÇÃO 2

NÚMERO DOS CANAIS	DENOMINAÇÃO DOS CANAIS
4	Canal 4 original do LANDSAT
9	Variação suavizada do canal 4 do LANDSAT
11	Variação suavizada do canal 6 do LANDSAT
12	Variação suavizada do canal 7 do LANDSAT

TABELA 5

ÍNDICES DE DESEMPENHO PARA ÁREAS DE TREINAMENTO

	CANAIS ORIGINAIS		J-M MÉDIA		J-M MÍNIMA		ENTROPIA	
L	5	6	5	6	5	6	5	6
Dm(%)	95,5	95,6	99,5	98,4	99,5	98,4	90,0	94,7
Am(%)	0,4	0,3	0,3	0,8	0,3	0,8	6,6	2,0
Cm(%)	4,2	4,2	0,3	0,9	0,3	0,9	3,2	3,3

TABELA 6

ÍNDICES DE DESEMPENHO PARA ÁREAS TESTE

	CANAIS ORIGINAIS		J-M MÉDIA		J-M MÍNIMA		ENTROPIA	
L	5	6	5	6	5	6	5	6
Dm(%)	78,0	80,6	81,1	83,8	81,1	83,8	77,8	91,9
Am(%)	4,8	0,3	13,1	6,6	13,1	6,6	22,0	6,6
Cm(%)	17,2	19,2	5,8	9,6	5,8	9,6	0,3	1,5

No entanto, os canais média borram regiões de transição e podem provocar erros de classificação em outras regiões não analisadas pela extração de parâmetros. Observe-se que os canais variação aparecem em ambos os casos, demonstrando a importância do uso da informação espacial para melhorar a precisão da classificação.

No futuro pretende-se testar o método em outras imagens com classes diferentes, a utilização conjugada de informação temporal, outros operadores de extração de atributos, e estudar a influência do método em classificação não-supervisionada.

6. REFERÊNCIAS BIBLIOGRÁFICAS

DUTRA, L.V. *Extração de atributos espaciais em imagens multiespectrais.* São José dos Campos, INPE, Fev. 1982. (INPE-2315-TDL/078).

II, F.A.M. *Seleção de atributos aplicada a imagens multiespectrais.* São José dos Campos, INPE, Jan. 1982. (INPE-2303-TDL/072).

SCHACHTER, B.J.; DAVIS, L.S.; ROSENFELD, A. Some experiments in image segmentation by clustering of local features values. *Pattern Recognition*, 11(1):19-28, Jan. 1979.

SWAIN, P.H.; KING, R.C. *Two effective feature selection criteria for multiespectral remote sensing.* West Lafayette, IN, Purdue University, 1973. (LARS Information Note 042673).

VELASCO, F.R.D.; PRADO, L.O.C.; SOUZA, R.C.M. *Sistema Maxver:* manual do usuário. São José dos Campos, INPE, Jul. 1978. (INPE-1315-NTI/110).

YOUNG, T.Y.; CALVERT, T.W. *Classification, estimation and pattern recognition.* New York, Elsevier, 1974.