1. Publicação nº	2. Versão	3. Data	5. Distribuição
INPE-3383-PRE/656		*Dez., 1984*	☐ Interna ☒ Externa ☐ Restrita
4. Origem: *DSC*	Programa: *PROCIM*		

6. Palavras chaves - selecionadas pelo(s) autor(es)

PROCESSAMENTO DE IMAGENS
SELEÇÃO ATRIBUTOS
ENTROPIA
DISTÂNCIAS ESTATÍSTICAS

7. C.D.U.: *621.376.5*

8. Título: *INPE-3383-PRE/656*

AVALIAÇÃO DA ENTROPIA GAUSSIANA E DA ENTROPIA DE SHANNON COMO CRITÉRIOS DE SELEÇÃO DE ATRIBUTOS

10. Páginas: *09*

11. Última página: *08*

12. Revisada por

Arry Carlos Buss Filho

13. Autorizada por

Nelson de Jesus Parada
Diretor Geral

9. Autoria *Luciano Vieira Dutra*
Fernando A. Mitsuo Ii
Nelson D.A. Mascarenhas

Assinatura responsável

14. Resumo/Notas *Esta pesquisa teve por objetivo avaliar o desempenho da entropia como um critério para seleção de atributos naturais ou derivados de imagens multiespectrais de recursos terrestres, particularmente da série LANDSAT. A precisão desejada de classificação de alvos terrestres utilizando imagens multiespectrais muitas vezes não é alcançada com o uso dos dados originais. Para aumentar a precisão pode-se utilizar a informação espacial contida nas imagens ou utilizar informação temporal corregistrando estas imagens com passagens de outras datas. Nestes casos, o número de atributos pode elevar-se consideravelmente, tornando a classificação uma tarefa computacionalmente pesada. Técnicas de seleção de atributos são utilizadas para escolher um menor subconjunto de atributos ou canais com uma dimensionalidade fixada, com o objetivo de maximizar a precisão de classificação, minimizando o esforço computacional. Um dos métodos utilizado como critério é a entropia que escolhe os canais que melhor representam os dados originais. Dentre as várias formulações para o cálculo da entropia, utilizaram-se a entropia gaussiana (baseada na hipótese de distribuição gaussiana para as classes) e a entropia da maneira definida por Shannon. Os experimentos demonstraram o bom desempenho do critério da entropia quando comparado aos métodos tradicionais (como o da distância Jeffreys-Matusita (J-M)). Embora a distância JM dê melhores resultados em alguns casos para áreas testes, a seleção de atributos usando a entropia gaussiana é bem mais rápida, recomendando-se portanto para a maioria dos casos. O uso de entropia pela definição de Shannon não traz vantagens adicionais, pois tem maior complexidade, embora não exija qualquer hipótese a priori a respeito da distribuição dos atributos das classes.*

15. Observações *Este trabalho será publicado nas anais da 4ª Reunião Anual da SELPER, Santiago, Chile, de 12 a 16 de Novembro de 1984.*

AVALIAÇÃO DA ENTROPIA GAUSSIANA E DA ENTROPIA DE SHANNON COMO CRITÉRIOS DE SELEÇÃO DE ATRIBUTOS

Luciano Vieira Dutra (1)

Fernando A. Mitsuo Ii (2)

Nelson D.A. Mascarenhas (3)

R E S U M O

Esta pesquisa teve por objetivo avaliar o desempenho da entropia como um critério para seleção de atributos naturais ou derivados de imagens multiespectrais de recursos terrestres, particularmente da série LANDSAT. A precisão desejada de classificação de alvos terrestres utilizando imagens multiespectrais muitas vezes não é alcançada com o uso dos dados originais. Para aumentar a precisão pode-se utilizar a informação espacial contida nas imagens ou utilizar informação temporal corregistrando estas imagens com passagens de outras datas. Nestes casos, o número de atributos pode elevar-se consideravelmente, tornando a classificação uma tarefa computacionalmente pesada. Técnicas de seleção de atributos são utilizadas para escolher um menor subconjunto de atributos ou canais com uma dimensionalidade fixada, com o objetivo de maximizar a precisão de classificação, minimizando o esforço computacional. Um dos métodos utilizado como critério é a entropia que escolhe os canais que melhor representam os dados originais. Dentre as várias formulações para o cálculo da entropia, utilizaram-se a entropia gaussiana (baseada na hipótese de distribuição gaussiana para as classes) e a entropia da maneira definida por Shannon. Os experimentos demonstraram o bom desempenho do critério da entropia quando comparado aos métodos tradicionais (como o da distância Jeffreys-Matusita (J-M)). Embora a distância JM dê melhores resultados em alguns casos para áreas testes, a seleção de atributos usando a entropia gaussiana é bem mais rápida, recomendando-se portanto para a maioria dos casos. O uso de entropia pela definição de Shannon não traz vantagens adicionais, pois tem maior complexidade, embora não exija qualquer hipótese a priori a respeito da distribuição dos atributos das classes.

---

(1) Engenheiro, Divisão de Suporte Computacional, Instituto de Pesquisas Espaciais (INPE), São José dos Campos, São Paulo, Brasil.
(2) Assistente de Pesquisas, Departamento de Processamento de Imagens, Instituto de Pesquisas Espaciais (INPE), São José dos Campos, São Paulo, Brasil.
(3) Pesquisador, Chefe de Departamento, Departamento de Processamento de Imagens, Instituto de Pesquisas Espaciais (INPE), São José dos Campos, São Paulo, Brasil.

A B S T R A C T

This research had the objective of evaluating the performance of the entropy as a criterion for selection of natural features or those derived from multispectral images of natural resources, particularly from the LANDSAT series. The desirable classification precision of land targets by means of multispectral images is often not attained with the original data. In order to improve the precision one can use spatial information in the images or temporal information by registering images from different dates. In these cases, the number of features can increase considerably, turning the classification into a computationally heavy process. Techniques for feature selection are used in order to choose a smaller subset of features or channels with a fixed dimension, with the objective of maximizing the classification precision and minimizing the computational effort. One of the used method is the entropy which selects the channels that best represent the original data. Among the several formulations for the computation of the entropy, the gaussian entropy (based on the hypothesis of gaussian distribution for the classes) and the entropy, according to Shannon's definition, were used. The experiments showed a good perform of the entropy, as compared to traditional methods (as the Jeffreys-Matusita (JM) distance). Although the JM distance provided better results in certain cases for test areas, feature selection by gaussian entropy is faster, therefore being recommended in general. The use of the entropy according to Shannon's definition does not bring additional advantages as it presents greater complexity, although it does not demand any a priori assumption about the distribution of classes attributes.

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 feitas a mão, impressões digitais, etc.

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

O extrator de atributos tem a função de reduzir os dados naturais medindo um certo conjunto de "atributos" ou "propriedades" que melhor caracterizem os objetos de interesse. Estes 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 determinísticas ou estatísticas.

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 divididas em técnicas de classificação paramétricas ou não-paramétricas.

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

Supondo 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 por um vetor N-dimensional $\underline{X} = [x_1, x_2, \ldots, x_N]$, ou por um ponto no espaço N-dimensional, chamado espaço característico Ω_X.

Normalmente, 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. 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. 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 a probabilidade do erro de classificação associado a cada subconjunto de K canais e, então, selecionando 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:

$$\left[\begin{matrix} N \\ n \end{matrix}\right] \triangleq \frac{N!}{n!(N-n)!}$$

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

$$\left[\begin{matrix} 12 \\ 4 \end{matrix}\right] = \frac{12!}{4!8!} = 495$$

integrações no espaço 4-dimensional. Mesmo em computadores muito rápidos, tais computações seriam proibitivas. Assim, métodos alternativos devem ser encontrados para a 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 têm sido propostas possui exatamente esta propriedade.

Contudo, diversas distâncias têm a característica de possuir limiares superior e/ou inferior para a probabilidade de erro associada a elas. 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 esta propriedade é subótima, pois não se minimiza, 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 esta 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 multiclasse.

A Distância B é função escalar das funções densidades de probabilidade de duas classes e é definida como:

$$B = -\ln \rho \text{ ou } \rho = e^{-B}, \tag{4}$$

onde

ρ = coeficiente de Bhattacharyya, dado por:

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

A distância JM é dada por:

$$d_{JM}^2 = 2(1-\rho) \longrightarrow d_{JM} = (2(1-\rho))^{1/2} \tag{6}$$

Para o caso de duas 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:

$$\frac{1}{4}\rho^2 \leq P_1P_1\rho^2 \leq \frac{1}{2}\left(1 - \sqrt{1 - 4P_1P_2\rho^2}\right) \leq P_E \leq \sqrt{P_1P_2}\ \rho \leq \frac{1}{2}\rho\ . \tag{7}$$

Para densidades gaussianas a Distância B é dada por:

$$B = \frac{1}{8}(\vec{\mu}_1 - \vec{\mu}_2)^T \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)(\vec{\mu}_1 - \vec{\mu}_2) + \frac{1}{2}\left\{\frac{\frac{1}{2}\,|\Sigma_1 + \Sigma_2|}{|\Sigma_1|^{1/2}\ |\Sigma_2|^{1/2}}\right\}, \tag{8}$$

onde

$\vec{\mu}_1$ e $\vec{\mu}_2$ são vetores-média, e Σ_1 e Σ_2 são matrizes de covariância para as classes 1 e 2, respectivamente.

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 têm duas classes, basta escolher o subconjunto com K atributos, para o qual a distância JM é maior. Para o caso de mais de duas 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 classe é maximizada. Outro critério é utilizado considerando o subconjunto que tenha a maior das distâncias JM mínima entre os pares de classes.

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

A 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:

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

onde

$|\Sigma|$ = determinante da matriz de covariância,

N = número de atributos.

A matriz de covariância pode ser calculada sobre o total das áreas de treinamento das classes envolvidas no processo de classificação, bastando para isto o cálculo da matriz de covariância NxN, onde N é o número total de atributos. Desta matriz se extraem todas as possíveis matrizes de dimensão kxk (k < N) e calcula-se a entropia gaussiana global para cada conjunto de k canais conforme a Expressão 9.

Pode-se também considerar cada classe separadamente e escolher o subconjunto de k atributos que maximize a soma das entropias gaussianas de cada classe, maximizando a Expressão 10:

$$S = \sum_{i=1}^{M} \ln |\Sigma_i| , \qquad (10).$$

onde

Σ_i = matriz de covariância da classe i

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

Outra maneira de calcular a entropia é a partir da própria definição da entropia de Shannon que é dada por:

$$H(x) = - \sum_{i=1}^{L} \rho_i \ln \rho_i , \qquad (11)$$

onde L é o número de pontos presentes na amostra, e ρ_i a frequência relativa de cada ponto na amostra.

De maneira similar à entropia gaussiana, a entropia de Shannon pode ser calculada sobre a amostra global das classes ou somadas às entropias de cada área de treinamento de cada classe para cada conjunto de k canais.

O cálculo da entropia de Shannon envolve maiores dificuldades, pois para o cálculo das frequências relativas para cada subconjunto de k canais é necessário extrair o histograma marginal k-dimensional a partir do histograma N-dimensional original, processo este muito custoso. Esta entropia tem a vantagem no entanto de não assumir nenhuma distribuição em particular.

3. RESULTADOS EXPERIMENTAIS

Os testes foram realizados com uma imagem TM (Thematic Mapper) da série LANDSAT com 6 canais, a qual cobre a área da Serra do Ramalho a sudoeste da Bahia, na margem esquerda do Rio São Francisco e a leste de Santa Maria da Vitória. O objetivo é escolher 4 canais, capacidade máxima de manipulação do imageador I100 pertencente ao Laboratório de Tratamento de Imagens Digitais do INPE em São José dos Campos. Foram selecionadas três classes bem distintas, dois conjuntos de áreas de treinamento e dois conjuntos de áreas testes com números diferentes de pontos nas áreas de treinamento e áreas testes para estudar a sensibilidade do método à variação no tamanho dessas áreas.

O segundo conjunto de áreas de teste contêm os pontos pertencentes às primeiras áreas teste mais os pontos que pertencem às primeiras áreas de treinamento e não pertencem às segundas áreas de treinamento. A Tabela 1 apresenta as classes e o número de pontos de cada área.

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).

A Tabela 2 apresenta os índices de desempenho para as áreas de treinamento e teste A. O índice de desempenho Dm é definido como a média da percentagem de classificação correta de cada área teste ou de treinamento, ponderada pelo número de pontos em cada área. Os parâmetros Am (Abstenção média) e Cm (Confusão média) têm definições similares. O limiar de classificação, que é um limiar aplicado ao log da função de verossimilhança do classificador, foi de 5 para todos os casos. Os resultados são apresentados para cada conjunto de canais escolhidos por cada critério. Como em todos os casos os canais escolhidos pela definição sobre a área global e os canais escolhidos pela soma da entropia de Shannon das áreas de cada classe foram os mesmos, apresenta-se apenas uma única coluna para os canais escolhidos pela definição.

Observa-se que os resultados foram similares para os canais escolhidos por todos os critérios, com uma ligeira vantagem para os canais escolhidos por distância JM.

Para avaliar a sensibilidade da escolha com relação à dimensão das áreas testes foi realizado um segundo conjunto de experiências com um classificador baseado nas áreas de treinamento B (Tabela 1).

A Tabela 3 apresenta os resultados para a classificação de áreas de treinamento B e teste A e B com o limiar de classificação 5. Observa-se que os resultados são ainda muito similares entre os diversos critérios de escolha de canais, embora com ligeira vantagem para os canais escolhidos por distância JM.

Os resultados confirmam a observação de que os canais escolhidos por cada critério sempre se colocam entre os primeiros selecionados para cada caso, conforme pode ser observado na Tabela 4, o que explica a pequena diferença entre os resultados para cada seleção.

Pode-se assim concluir que, no caso onde foram utilizadas classes bem separadas e classificadores estabelecidos com base em áreas testes com número de pontos maior que 25n, sendo n=4 o número de canais utilizados, basta utilizar o critério de seleção por entropia gaussiana global, o mais rápido de todos os critérios.

4. CONCLUSÕES FINAIS E FUTUROS DESENVOLVIMENTOS

Para as condições em que foram realizados os testes, os diversos critérios não demonstraram diferenças sensíveis de resultados; no entanto mudanças nas condições, tais como uso de áreas de treinamento da ordem de 10n (n é o número de canais), escolha de classes não tão distintas quanto àquelas utilizadas aqui e uso de classes fortemente não-gaussianas como água, por exemplo, podem alterar o resultado.

A entropia no entanto já demonstrou um forte potencial para ser utilizado em problemas de seleção de atributos.

Como trabalhos futuros sugerem-se os testes em outras condições e a utilização de classes de relevo em vez de classes de uso do solo.

5. BIBLIOGRAFIA

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

- Swain, P.H. and King, R.C. "Two effective feature selection criteria for multispectral remote sensing". West Lafayette, IN, Purdue University, 1973. (LARS Information Note 042673).

- Velasco, F.R.D.; Prado, L.O.C. e Souza, R.C.M. "Sistema Maxver: manual do usuário". São José dos Campos, INPE, jul. 1978. (INPE-1315-NTI/110).

- Young, T.Y. and Calvert, T.W. "Classification, estimation and pattern recognition". New York, Elsevier, 1974.

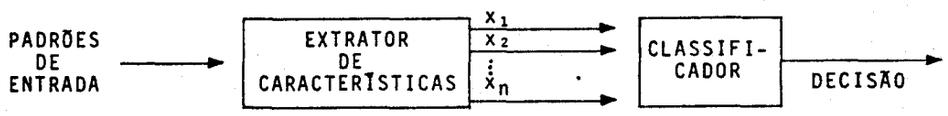

FIGURA 1. SISTEMA DE CLASSIFICAÇÃO DE PADRÕES

CLASSES	ÁREAS DE TREINAMENTO		ÁREAS TESTES	
	A	B	A	B
Arenito	300	100	200	400
Calcáreo	252	120	180	312
Uso-solo	300	120	180	360

TABELA 1 - CLASSES USADAS

CRITÉRIO DE SELEÇÃO	JM MÉDIA		SOMA ENTROPIAS GAUSSIANAS		GAUSSIANA GLOBAL		DEFINIÇÃO (SHANNON)	
CANAIS ESCOLHIDOS	3457		1457		1345		1345	
	TREIN.	TESTE	TREIN.	TESTE	TREIN.	TESTE	TREIN.	TESTE
Dm (%)	99,8	85,4	99,8	84,8	99,5	84,8	99,5	84,8
Am (%)	0,2	14,6	0,2	15,2	0,5	15,2	0,5	15,2
Cm (%)	-	-	-	-	-	-	-	-

TABELA 2 - ÍNDICES DE DESEMPENHO PARA ÁREAS DE TREINAMENTO A E ÁREAS TESTES A

CRITÉRIO DE SELEÇÃO	JM MÉDIA			SOMA ENTROPIAS GAUSSIANAS			GAUSSIANA GLOBAL			ENTROPIA PELA DEFINIÇÃO		
CANAIS ESCOLHIDOS	2457			1457			1345			3457		
	TREIN. B	TESTE A	TESTE B	TREIN. B	TESTE A	TESTE B	TREIN. B.	TESTE A	TESTE B	TREIN. B	TESTE A	TESTE B
Dm (%)	99,7	83,0	89,1	99,4	82,3	88,5	99,1	81,8	88,4	99,7	81,8	88,4
Am (%)	0,3	17,0	10,9	0,6	17,7	11,5	0,9	18,2	11,6	0,3	18,2	11,6
Cm (%)	-	-	-	-	-	-	-	-	-	-	-	-

TABELA 3 - ÍNDICES DE DESEMPENHO PARA ÁREAS DE TREINAMENTO B E ÁREAS TESTES A E B

CRITÉRIO DE SELEÇÃO / ORDEM DE ESCOLHA	JM MÉDIA	SOMA ENTROPIAS GAUSSIANAS	GAUSSIANA GLOBAL	ENTROPIA PELA DEFINIÇÃO
1º	2457	1457	1345	3457
2º	2345	1345	3457	1245
3º	3457	3457	1457	1345

TABELA 4 - CANAIS ESCOLHIDOS POR CADA CRITÉRIO USANDO ÁREAS DE TREINAMENTO B.