SOBRE O USO DA MINIMIZAÇÃO DA ASSIMETRIA COMO CRITÉRIO DA VALIDAÇÃO DE AGREGAMENTO

Luciano Vieira Dutra
José Carlos Moreira
Instituto de Pesquisas Espaciais - INPE
Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq
Caixa Postal 515 - 12200 - São José dos Campos - SP - Brasil

RESUMO

Um dos principais problemas apresentados quando se utilizam métodos de classificação não-supervisionada ("clustering" ou agregamento) é a avaliação do resultado, ou seja, determinar se os agregamentos resultantes são significativos e correspondem a alvos existentes na natureza. Uma condição normalmente aceita estabelece que os atributos que descrevem os alvos naturais seguem uma distribuição gaussiana, ou unimodal simétrica em torno da média. Como o momento de terceira ordem (assimetria) de uma distribuição simétrica é zero, pode-se utilizar deste fato para determinar se um conjunto de agregados resultantes de uma classificação é composto de distribuições simétricas. O método proposto consiste em subdividir a massa de dados em diversos números de classes segundo o algoritmo de K-médias e aplicar o teste de minimização da assimetria média das distribuições das classes para escolher uma dada subdivisão. A avaliação desta subdivisão, com a finalidade de testar o método, foi realizada sobre diversas massas de dados com o número de classes previamente conhecido. Sobre esta massa de dados foram aplicados também outros testes de validação conhecidos. Os resultados demonstraram que este critério conduziu a divisões muito próximas da realidade, com desempenho superior aos outros critérios utilizados.

ABSTRACT

One of the main problems with the application of clustering algorithm is the evaluation of the results in order to decide if the clusters correspond to real targets (this is known as validity studies). A well-known condition stablish that features describing natural targets follow a Gaussian or at least a symetrycal and unimodal distribution. The third momentum (asymmetry) of a symmetrical distribution is zero. So, by calculating this momentum for all clusters in a given partition, one can determine if this partition is composed of symmetrical distributions. The proposed method consists of partitioning

the data in several number of classes using the well-known K-mean algorithm, and choosing the partition that yelds the minimum mean asymmetry for all partitions. The method was tested with several groups of data, each one containing a certain number of classes known a priori. Other known validity methods were applied for the same groups of data. The results demonstrate that the mean asymmetry criterion led to partitions that in most cases were near the reality, with a better performance than the other methods that were tested.

1. INTRODUÇÃO

Desde de o advento dos computadores digitais tem havido um constante esforço no sentido de desenvolver métodos automáticos de decisão em tarefas que efetuadas manualmente seriam monótonas, repetitivas e menos precisas. A fotointerpretação é uma destas tarefas.

O uso de imagens de satélites de recursos terrestres veio facilitar a tarefa de analisar, com rapidez e economia, grandes áreas agrícolas, possibilitando o levantamento de dados tais como áreas cultivadas e previsão de safras.

Os métodos computarizados de partição (classificação) de imagens multiespectrais em regiões, os quais associam a cada região uma classe ou cultura, foram desenvolvidos com base na teoria estabelecida para a resolução dos problemas gerais de classificação de padrões e análise estatística.

A classificação de imagens digitais pode ser feita de modo *supervisionado*, ou seja, o pesquisador fornece ao computador as características dos alvos que pretende identificar na imagem utilizando-se de *áreas de treinamento* e o algoritmo particiona a cena em regiões relacionadas. De *modo não-supervionado*, quando o algoritmo procura automaticamente particionar imagens segundo relações intrínsecas aos dados. Neste último caso os algoritmos têm o nome genérico de "clustering" ou agregamento. Os algoritmos de agregamento têm recebido grande atenção na literatura em diversos campos do conhecimento e grande parte de sua formulação teórica ainda está para ser estabelecida.

Um dos principais problemas apresentados é o de descobrir o número correto de classes ou agregados existentes em uma massa de dados e se estes agregados têm significado real ou não (Dubes et alii, 1979).

Este trabalho desenvolve um método para escolher uma dentre as várias participações produzidas pelo bem conhecido algoritmo das K-médias baseado em certas características desejadas para os agregados e compara este método com outras existentes na literatura. O problema da decisão se os agregados resultantes têm ou não significado é deixado ao pesquisador.

2. ALGORITMOS DE AGREGAMENTO

Uma imagem digital multiespectral é uma coleção de K matrizes com m linhas e n colunas. Cada elemento de cada matriz é a média da reflectância do alvo em uma área, denominada resolução, e em uma determinada banda espectral. Um conjunto de elementos da mesma localização que identifica um ponto no terreno é denominado "pixel".

Cada "pixel" define portanto um ponto no espaço K-dimensional de medidas (atributos), onde o valor de cada ordenada no n-ésimo eixo é o valor do elemento na n-ésima matriz do canal da imagem.

Pode-se representar uma imagem qualquer no espaço de atributos plotando cada "pixel" desta imagem neste espaço. Eventualmente, vários pontos localizar-se-ão no mesmo local no espaço de atributos. Caso se conte o número de superposições neste espaço, obter-se-á o histograma K-dimensional desta imagem. O objetivo do uso dos algoritmos de agregamento é localizar nuvens densas de pontos neste espaço.

A densidade é função da quantidade de pontos próximos uns dos outros, além do número de vezes que cada um deles se repete (chamado população do ponto).

Um algoritmo clássico de partição de um certo histograma K-dimensional, que representa uma imagem, é o chamado algoritmo das "K-médias" que pode ser assim explicado (Tou et alii, 1974):

1) escolhe-se da massa de dados um número L de centros iniciais para dividir a imagem em L classes;

2) associa-se cada ponto do histograma ao centro que tiver menor distância;

3) calculam-se novos centros para as classes usando as coordenadas dos pontos de cada classe e a média ponderada pela população;

4) termina-se a partição se os novos centros forem diferentes dos anteriores por uma distância menor que um certo limiar, senão volta-se ao passo 2.

Para verificar qual o número ótimo de partições de uma certa massa de dados, aplica-se o algoritmo de K-médias para vários números de centros iniciais, normalmente de 1 a K-classes, e para cada partição calcula-se uma série de parâmetros para avaliar a validade desse conjunto particionado. A partição em $\overline{K}$ classes é selecionada com base nestes parâmetros, quando passam por um máximo ou um mínimo, dependendo dos parâmetros.

Os problemas básicos que podem aparecer na aplicação desse algoritmo são os seguintes:

1) Escolha dos centros iniciais: Neste caso são escolhidos os L pontos mais populosos que distem dentre si uma distância maior que $\emptyset$. Outras escolhas podem ser aplicadas.

2) Eventualmente o algoritmo pode não convergir para uma certa precisão requerida no passo 4. Neste caso deve-se relaxar a precisão.

Esses parâmetros mencionados são importantes a medida que alteram, em certos casos, os resultados para cada número de classes desejadas.

3. ÍNDICES DE VALIDAÇÃO

Para o tipo de algoritmo de agregamento mencionado, o principal problema é determinar o número correto de agregamentos. Um método possível de obter uma medida da qualidade do agregamento é representado por um parâmetro "Beta" (Coleman and Andrews, 1979), função de medidas sobre a matriz de espalhamento intra-agregados e a matriz de espalhamento entre-agregados.

A matriz de espalhamento intra-agregados é baseada no espalhamento dos dados em torno das médias e é dada por:

$$S_{\omega} = \frac{1}{K} \sum_{k=1}^{K} \varepsilon\{(\vec{x} - \vec{\mu}_k)\ (\vec{x} - \vec{\mu}_k)^t\}\ , \tag{3.1}$$

onde x é vetor de atributos e

$$\varepsilon\{\cdot\} = \frac{1}{M_k} \sum_{x_i \varepsilon S_k} (\vec{x}_i - \vec{\mu}_k)\ (\vec{x}_i - \vec{\mu}_k)^t\ , \tag{3.2}$$

onde:

$\vec{\mu}_k$ - média do k-ésimo agregado,

M_k - número de elementos no k-ésimo agregado,

$\vec{x}_i$ - um elemento no k-ésimo agregado,

K - número total de agregados.

A matriz de espalhamento entre-agregados pode ser definida para $K \geq 2$ por:

$$S_b = \frac{1}{K} \sum_{k=1}^{K} (\vec{\mu}_k - \vec{\mu}_0)\ (\vec{\mu}_k - \vec{\mu}_0)^t\ , \tag{3.3}$$

onde μ_0 é a média total da mistura das classes dada por:

$$\mu_0 = \frac{1}{M} \sum_{i=1}^{M} \vec{x}_i\ . \tag{3.4}$$

O objetivo de usar essas matrizes de espalhamento é medir a separabilidade dos agregados. Para isto, derivam-se certos parâmetros denominados "Beta":

$\beta_1 = tr(S_w^{-1} S_b)$,

$\beta_2 = \ln \{|S_w + S_b|/|S_w|\}$,

$$\beta_3 = \text{tr } S_b - \mu(\text{tr } S_w - c), \tag{3.5}$$

$$\beta_4 = \text{tr } S_b / \text{tr } S_w,$$

$$\beta_5 = \text{tr } S_b \cdot \text{tr } S_w,$$

onde tr(·) é traço de uma matriz. Quando β_3 é usado, o procedimento é maximizar tr Sb sujeito à restrição tr Sw = c.

O uso dos parâmetros "Beta" requer que um "joelho" no gráfico "Beta" versus número de classes seja detectado.

É possível que vários "joelhos" venham a ser detectados.

O presente trabalho sugere a utilização do momento de terceira ordem de uma distribuição (assimetria) como medida de validade de uma certa divisão. O objetivo é detectar a divisão que isola os agregados que formem nuvens com densidade simétrica em torno da média de cada agregado, decorrente da hipótese geralmente aceita de que os dados naturais seguem uma distribuição gaussiana, ou pelo menos unimodal e simétrica.

Para isso, calculam-se as assimetrias marginais de cada agregado, obtendo-se em seguida um índice a ser minimizado, representado pela média de todas as assimetrias marginais.

A assimetria para o agregado j, canal K, é dada por (Gnanadesikan, 1977):

$$a_{jk} = \sqrt{b_1} = \frac{\sqrt{n} \sum_{i=1}^{n} (x_i - \bar{x})^3}{\left\{ \sum_{i=1}^{n} (x_i - \bar{x})^2 \right\}^{3/2}}, \tag{3.6}$$

onde n é o número de pontos do agregado j e $\bar{x}$ é a média do agregado j, canal k.

Portanto, basta minimizar a quantidade $\left(\frac{\sum_j \sum_k a_{jk}}{N} \right)$ para que uma determinada divisão em N classes seja escolhida.

4. RESULTADOS EXPERIMENTAIS

Dois tipos de conjuntos de dados quadridimensionais foram utilizados. O primeiro é composto de dados gaussianos artificiais e o outro contém dados de imagens multiespectrais do LANDSAT 3, órbita 234, ponto 26, de 22 de julho de 1978 (região do Município Ribas do Rio Pardo, MS, fazenda Mutum da Cia. Itapeva Florestal).

Três conjuntos de dados artificiais foram obtidos com 2, 3 e 4 distribuições gaussianas, com cerca de 250-300 pontos em cada distribuição. Os dados gerados são misturados e ordenados posteriormente pela população.

Para cada arquivo de dados aplica-se o algoritmo das K-médias sucessivamente de 1 a 7 classes e calculam-se os parâmetros β e assimetria.

O arquivo com 2 distribuições gaussianas foi gerado com base nos seguintes parâmetros:

$$\vec{m}_1 = (70, 70, 40, 40), \quad \Sigma_1 = \begin{pmatrix} 25 & 0 & 0 & 0 \\ 0 & 25 & 0 & 0 \\ 0 & 0 & 25 & 0 \\ 0 & 0 & 0 & 25 \end{pmatrix} \tag{4.1}$$

$$\vec{m}_2 = (40, 40, 70, 70), \quad \Sigma_2 = \begin{pmatrix} 25 & 0 & 0 & 0 \\ 0 & 25 & 0 & 0 \\ 0 & 0 & 25 & 0 \\ 0 & 0 & 0 & 25 \end{pmatrix} \tag{4.2}$$

Ambas as classes têm 297 pontos cada uma.

O arquivo com 3 classes foi estabelecido com base nos dados:

$$\vec{m}_1 = (70, 70, 40, 40), \quad \Sigma_1 = \begin{pmatrix} 25 & 0 & 0 & 0 \\ 0 & 25 & 0 & 0 \\ 0 & 0 & 25 & 0 \\ 0 & 0 & 0 & 25 \end{pmatrix} \tag{4.3}$$

$$\vec{m}_2 = (40, 40, 70, 70), \quad \Sigma_2 = \begin{pmatrix} 25 & 0 & 0 & 0 \\ 0 & 25 & 0 & 0 \\ 0 & 0 & 25 & 0 \\ 0 & 0 & 0 & 25 \end{pmatrix} \tag{4.4}$$

$$\vec{m}_3 = (70, 40, 70, 40), \quad \Sigma_3 = \begin{pmatrix} 25 & 0 & 0 & 0 \\ 0 & 25 & 0 & 0 \\ 0 & 0 & 25 & 0 \\ 0 & 0 & 0 & 25 \end{pmatrix} \tag{4.5}$$

O arquivo com 4 classes foi estabelecido com base nos dados seguintes:

$$\vec{m}_1 = (30, 80, 40, 70), \quad \Sigma_1 = \begin{pmatrix} 25 & 0 & 0 & 0 \\ 0 & 36 & 0 & 0 \\ 0 & 0 & 16 & 0 \\ 0 & 0 & 0 & 36 \end{pmatrix} \tag{4.6}$$

$$\vec{m}_2 = (80, 40, 70, 30), \quad \Sigma_2 = \begin{pmatrix} 25 & 0 & 0 & 0 \\ 0 & 25 & 0 & 0 \\ 0 & 0 & 36 & 0 \\ 0 & 0 & 0 & 16 \end{pmatrix} \tag{4.7}$$

$$\vec{m}_3 = (70,\ 30,\ 80,\ 40), \quad \Sigma_2 = \begin{pmatrix} 25 & 0 & 0 & 0 \\ 0 & 25 & 0 & 0 \\ 0 & 0 & 36 & 0 \\ 0 & 0 & 0 & 25 \end{pmatrix} \tag{4.8}$$

$$m_4 = (40,\ 70,\ 30,\ 80), \quad \Sigma_3 = \begin{pmatrix} 16 & 0 & 0 & 0 \\ 0 & 36 & 0 & 0 \\ 0 & 0 & 16 & 0 \\ 0 & 0 & 0 & 36 \end{pmatrix} \tag{4.9}$$

A Tabela 1 resume o resultado do processamento para esses 3 arquivos relativos aos parâmetros β_2, β_5 e assimetria.

Pode-se observar que os parâmetros β_2 e β_5 não apresentam nenhum "joelho", sendo que a assimetria passou por um mínimo no número correto de classes.

Os casos em que a assimetria não foi zero, no número correto de classes, devem-se ao fato de o algoritmo de agregamento classificar como pertencentes a um agregado pontos gerados para outro agregado gaussiano.

Para o experimento prático foram gerados arquivos com dados de 2, 3, 4 e 5 classes.

As classes utilizadas foram as seguintes:

1. *Eucalipto*,
2. *Eucalipto novo*,
3. *Solo preparado*,
4. *Solo exposto*,
5. *Cerrado*.

Foram preparados 4 arquivos de dados, cada qual com um número diferente de classes, de acordo com a distribuição abaixo:

- arquivo 1 (2 classes) classes: 1 e 2,
- arquivo 2 (3 classes) classes: 1, 3 e 5,
- arquivo 3 (4 classes) classes: 1, 2, 3 e 4,
- arquivo 4 (5 classes) classes: 1 a 5.

Os resultados encontram-se sumariados na Tabela 2.

Dos resultados apresentados pode-se observar a presença de mínimos locais no número correto de classes para cada caso, embora para 4 e 5 classes o mínimo geral se deu em 1 classe. Isto se deve ao fato de que quanto mais generalizado se forma um arquivo, ou seja, com um maior número de classes representadas, os dados tendem a se espalhar mais uniformemente, obtendo-se assim uma simetria bem maior em torno da média global. Este resultado é confirmado ao observar-se que o histograma de uma imagem total composta de temas não muito díspares tende a apresentar histogramas simétricos.

TABELA 1

PARÂMETROS DE VALIDAÇÃO PARA DADOS GAUSSIANOS

Nº DE CLASSES EXISTENTES	Nº DE CLASSES DIVIDIDAS	1	2	3	4	5	6	7
2	assimetria	0,283	0,0127	0,779	1,196	1,350	1,190	0,716
	β_5 ($x10^3$)	-	450	307	243	211	181	160
	β_2	-	3,382	3,912	4,379	4,144	4,196	4,300
3	assimetria	2,375	0,196	10^{-5}	0,055	0,141	0,226	0,626
	β_5 ($x10^3$)	-	202	79,1	75,6	74,8	70,8	63,67
	β_2	-	3,03	5,43	5,77	6,127	6,398	6,644
4	assimetria	8,625	0,303	0,185	0,110	0,193	0,459	0,565
	β_5 ($x10^3$)	-	1,608	1,602	1,598	1,348	1,284	1,186
	β_2	-	1,733	2,476	2,679	2,826	3,465	3,667

TABELA 2

PARÂMETROS DE VALIDAÇÃO PARA DADOS TEMÁTICOS

* escolha cuidadosa das classes

Nº DE CLASSES EXISTENTES	Nº DE CLASSES DIVIDIDAS	1	2	3	4	5	6	7
2 (arquivo 1)	assimetria	0,8854	0,874*	2,303	1,028	1,22	-	-
	β_5	-	2,739	2,195	2,086	1,878	-	-
	β_2	-	1,909	2,865	3,293	3,811	-	-
3 (arquivo 2)	assimetria	2,317	4,158	1,407*	1,443	1,589	1,421	-
	β_5	-	34,348	20,006	15,641	12,959	9,742	-
	β_2	-	2,128	4,07	4,67	5,039	5,899	-
4 (arquivo 3)	assimetria	0,5412	1,920	1,850	1,331*	1,360	1,389	-
	β_5	-	11,665	8,435	8.386	7.404	6.011	-
	β_2	-	1,53	2,794	3,582	4,594	5,032	-
5 (arquivo 5)	assimetria	0,6426	1,319	2,376	2,276	1,920*	1,964	2,166
	β_5	-	15,271	12,924	9,441	7,597	6,708	5.404
	β_2	-	1,666	2,540	3,709	4,565	4,933	4,915

A análise visual do resultado da classificação indica razoável consonância com classes realmente existentes, embora nos casos em que se trabalhou com um menor número de classes, como por exemplo no caso de análise com 3 classes, uma das classes classificadas englobava todas as outras na imagem.

Para testar a sensibilidade do teste com relação à junção com classes próximas, foram preparados mais 2 arquivos com 3 classes:

- arquivo 5 (3 classes): 2, 4 e 5,
- arquivo 6 (3 classes): 1, 3 e 4,

onde há confusão entre classes *eucalipto novo* e *solo exposto* e confusão entre *solo preparado* e *solo exposto*. Os resultados encontram-se sumariados na Tabela 3 através da qual pode-se observar que não foi bom o desempenho de nenhum parâmetro para os arquivos 5 e 6.

TABELA 3

ÍNDICES DE VALIDAÇÃO PARA MISTURAS COM 3 CLASSES NATURAIS

Nº DE CLASSES EXISTENTES	Nº DE CLASSES DIVIDIDAS	1	2	3	4	5	6	7
3 (arquivo 2)	assimetria	2,317	4,158	1,407*	1,443	1,589	1,421	
	β_5	-	34,348	20,006	15,641	12,959	9,742	
	β_2	-	2,128	4,07	4,67	5,039	5,899	
3 (arquivo 5)	assimetria	1,943*	3,719	5,266	4,865	4,130	3,50	
	β_5	-	10,900	6,500	4.780	5,770	6,110	
	β_2	-	1,638	2,971	3,059	3,716	4,080	
3 (arquivo 6)	assimetria	1,746	2,786	1,745	1,711	1,228*	1,543	
	β_5	-	16,950	15,260	11,216	10,292	7,820	
	β_2	-	1,89	3,221	3,848	4,569	4,538	

* mínimos

5. CONCLUSÃO

Os parâmetros "Beta" não obtiveram melhor desempenho, não apresentando "joelhos" que determinassem o número de classes presentes nos dados; ao contrário da assimetria que, embora possa dar resultado errôneo nos casos de misturas com classes muito semelhantes, apresentou resultado satisfatório em muitos casos.

Os resultados são também muito sensíveis aos parâmetros aplicados ao algoritmo de agregamento, tais como precisão de convergência, número máximo de iterações, método de escolha dos centros iniciais, etc.

Conclui-se finalmente que o método apresentado, embora promissor, não é muito eficaz, sugerindo ainda pesquisas exaustivas. Como sugestão propõe-se a análise em conjunto com momentos de quarta ordem (curtose) e testes para avaliar a normalidade n-dimensional dos dados das classes. Sugere-se também que, para cada divisão em K classes, deve-se procurar distinguir nessas K classes um número menor de classes que atenda às premissas de um bom agregado.

6. REFERÊNCIAS BIBLIOGRÁFICAS

COLEMAN; G.B.; ANDREWS, H.C. Image segmentation by clustering. *Proceedings of IEEE*, 67(5):773-785, May, 1979

DUBES, R.; JAIN, A.K. Validity studies in clustering methodologies. *Pattern Recognition*, 11:235-254, 1979.

GNANADESIKAN, R. *Methods for statistical data analysis of multivariate observations*, New York, N.Y., John Wiley, 1977.

TOU, J.; GONZALES, R.C. *Pattern recognition principles*. Reading, Mass, Addison-Wesley, 1974.