Anais XIII Simpósio Brasileiro de Sensoriamento Remoto, Florianópolis, Brasil, 21-26 abril 2007, INPE, p. 6143-6150.

Seleção de atributos usando algoritmos genéticos para classificação de regiões

Joelma Carla Santos
João Ricardo de Freitas Oliveira
Luciano Vieira Dutra
Sidnei João Siqueira Sant'Anna
Camilo Daleles Rennó

Instituto Nacional de Pesquisas Espaciais - INPE
Caixa Postal 515 - 12245-970 - São José dos Campos - SP, Brasil
{joelma, joao, dutra, sidnei, camilo}@dpi.inpe.br

Abstract. Good feature extraction and selection methods are crucial for adequate performance of remote sensing data classification methods. In general, a big set of features causes degradation in results of classification. When data dimensionality is very high, a search strategy should be used to select the subset of features that gives largest separability between classes. But good results of classification also are dependents of classifier. In this paper different chromosome encoding form are compared. The performed tests showed that the performance of permutation encoding was better than binary encoding. Two images were used: a synthetic image and quickbird image with 0.6 m of resolution. The classification was done with Euclidian distance and good values of kappa coefficient were reached.

Palavras-chave: remote sensing, image processing, genetic algorithm, feature extraction, feature selection, region classification, sensoriamento remoto, processamento de imagens, algoritmo genético, extração de atributos, seleção de atributos, classificação de regiões.

1. Introdução

Os avanços tecnológicos cada vez maiores na área de sensores remotos têm gerado imagens com maior poder de discriminação de alvos da superfície terrestre. Com a possibilidade de discriminação de alvos urbanos, torna-se maior o número de aplicações dos dados de sensoriamento remoto para estudos relativos ao sistema urbano (Donnay et al., 2001) e aumenta-se a precisão das informações obtidas a partir deles (Souza et al, 2003).

Para tratamento e obtenção de informações das imagens são usados sistemas de processamento de imagens que abrangem operações como aquisição, armazenamento, processamento e visualização. A operação de processamento pode envolver várias etapas como pré-processamento, segmentação, extração de atributos, treinamento, seleção de atributos e classificação (Marques Filho e Vieira Neto 1999). O foco deste trabalho é direcionado às etapas de extração e seleção de atributos.

Uma grande dimensionalidade do espaço de atributos pode causar degradação na classificação e um alto custo de processamento. O objetivo dos algoritmos de seleção de atributos é escolher o menor subconjunto que ofereça a melhor classificação em conformidade com custos computacionais razoáveis.

O método de classificação utilizado neste trabalho é a classificação por regiões. Estes classificadores utilizam, além da informação espectral de cada pixel, a informação espacial que envolve a relação entre os pixels e seus vizinhos. A grande vantagem de classificar regiões é que o número de variáveis consideradas na classificação pode aumentar, pois, além dos dados espectrais, é possível descrever cada região usando parâmetros de forma, textura e associação entre objetos.

O objetivo deste trabalho é aplicar métodos de extração e seleção de atributos para classificação de regiões. Atributos de textura e forma são extraídos das regiões em estudo e métodos de seleção são aplicados a fim de reduzir a dimensionalidade do espaço de atributos.

2. Extração de Atributos

O objetivo da extração de atributos é caracterizar medidas associadas ao objeto que se deseja extrair, de forma que as medidas sejam similares para objetos similares e diferentes para objetos distintos (Duda et al., 2001). O processo de extração de atributos é fundamental para a obtenção de bons resultados na classificação, pois nesta fase é criado o conjunto de características usado para realização de treinamento e classificação.

O sistema de processamento de imagens *Texture 2.0* (Rennó et al., 1998) fornece um ambiente amigável ao usuário para extração e análise de medidas texturais de imagens, e realiza a classificação de regiões baseada na pré-seleção de medidas. No trabalho desenvolvido por Oliveira et al. (2005) foram acrescentadas ao sistema funções de extração e análise de medidas de formas tais como área, perímetro, complexidade e circularidade além de alguns algoritmos para seleção de atributos. No desenvolvimento deste trabalho também foram incluídos ao sistema os atributos de forma: deficiência convexa, assimetria, densidade, direção principal e comprimento de fibra.

Os atributos de forma são ferramentas poderosas para a discriminação de objetos que possuem a mesma aparência espectral (Andrade et al., 2002). A análise da forma do objeto é de fundamental importância para estudos relacionados com o espaço urbano. Construções e asfalto na maioria das vezes apresentam aparência espectral muito similar, o que dificulta a discriminação entre esses objetos. No entanto essa discriminação se torna mais fácil se for levado em consideração que construções possuem uma forma retangular, ao contrário do asfalto que possui uma forma mais alongada.

As texturas contêm informações importantes sobre o arranjo estrutural das superfícies e seus relacionamentos com o ambiente ao redor. Mais informações sobre os atributos de textura existentes no sistema Texture em Oliveira et al. (2005).

3. Seleção de Atributos

A seleção de atributos pode ser vista como um processo de busca onde o algoritmo usado deve encontrar o menor subconjunto de atributos com a melhor acurácia de classificação (Pappa at al. 2002b). Em um maior nível de abstração, a seleção de atributos pode ser dividida em duas partes: o método de busca e a função de aptidão usada para medir a qualidade dos subconjuntos de atributos (Pappa et al., 2002a). Os algoritmos de seleção de atributos são divididos em três grupos: exponenciais (busca exaustiva), seqüenciais e randômicos (algoritmo genético) (Boz, 2002). O algoritmo genético, método de busca utilizado neste trabalho, será detalhado na seção 3.1.

O processo de avaliação da qualidade dos subconjuntos de atributos é executado com base em distâncias estatísticas entre pares de classes. Uma das formas de medir a distância entre classes no espaço de atributos é através da distância de Bhattacharyya.

$$B_{ki} = \frac{1}{8}(\mu_k - \mu_i)^t \left[\frac{\Sigma_k + \Sigma_i}{2}\right]^{-1} (\mu_k - \mu_i) + \frac{1}{2}\ln \frac{\left|\frac{\Sigma_k + \Sigma_i}{2}\right|}{\sqrt{|\Sigma_k||\Sigma_i|}}$$

onde Σ_k e Σ_i são as matrizes de covariância das classes *k* e *i*, μ_k e μ_i são os respectivos vetores de média (Bhattacharyya, 1943)..

Para selecionar atributos com base em distâncias entre classes, é necessário definir uma função critério que possa avaliar a separabilidade entre todas as classes de uma maneira global. Para isto pode-se usar, por exemplo, uma simples operação de soma, média ou desvio padrão. Visando otimizar o conjunto de atributos para minimizar a probabilidade de erro na

classificação, deve-se escolher uma função critério que apresente valores maiores quanto maior a facilidade de discriminação entre os padrões de classes diferentes. A função critério usada neste trabalho foi a distância média Jefferys-Matusita (JM). A distância *JM* entre as classes *k* e *i* é dada por:

$$JM_{ki} = \sqrt{2(1-e^{-B_{ki}})} \qquad JM \in \left[0, \sqrt{2}\right]$$

onde B_{ki} representa a distância de Bhattacharyya entre as classes *k* e *i*.

3.1 Algoritmos Genéticos

Os algoritmos genéticos (AGs), desenvolvidos por John Holland, são métodos de busca e otimização baseados nos mecanismos de evolução dos seres vivos (Goldberg 1989). Estes algoritmos baseiam-se na teoria do naturalista Charles Darwin (1859), que afirma que os indivíduos mais adaptados ao seu ambiente são os que possuem maior chance de sobreviver e gerar descendentes.

O primeiro passo de um AG é a geração de uma população inicial de indivíduos caracterizados por seus cromossomos que codificam as possíveis soluções do problema. Durante o processo evolutivo, esta população é avaliada e cada cromossomo recebe um valor, calculado pela função de aptidão, refletindo sua habilidade de adaptação a determinado ambiente. Os cromossomos mais aptos são selecionados e os menos aptos são descartados (Darwinismo). Os indivíduos selecionados sofrem cruzamentos e mutações, gerando descendentes para a população da próxima geração. Este processo é repetido até que uma solução satisfatória seja encontrada (Goldberg 1989; Lacerda e Carvalho 1999).

Para selecionar os indivíduos foi utilizado o método da roleta, que segundo Goldberg (1989), cada indivíduo possui uma fatia da roleta proporcional à sua adaptação. A cada giro da roleta um indivíduo é selecionado, tendo maior chance aqueles que possuem as maiores fatias.

O cruzamento consiste basicamente em misturar o material genético de dois indivíduos da população, produzindo dois novos indivíduos (filhos) que herdam características dos pais.

A operação de mutação evita a convergência prematura do algoritmo, introduzindo na busca novas regiões do espaço de soluções (Oliveira, 1998).

A função de aptidão é utilizada para determinar a qualidade de uma solução candidata (Pappa 2002a). Ela oferece ao AG uma medida da aptidão de cada indivíduo da população (Goldberg 1989). A escolha de uma função de aptidão apropriada é um passo essencial para o sucesso de uma aplicação de AG (Vafaie e De Jong, 1992, 1993; Oliveira, 1998).

Neste trabalho foram usados dois algoritmos genéticos diferentes: AG binário e AG permutação.

No AG binário o genótipo de um cromossomo é representado por um vetor binário onde cada elemento do vetor é um dígito binário (0 ou 1); no caso deste trabalho, 1 representa a presença e 0 a ausência do atributo associado. Para esta codificação foi usado o cruzamento de um ponto (Goldberg, 1989).

Para o AG permutação a representação do cromossomo é feita através de uma lista de elementos que representam todos os atributos. Cada permutação em uma lista de elementos gera um cromossomo diferente. Permutações de *n* elementos são seqüências desses *n* elementos sem que nenhum seja repetido. Para esta codificação foi usado o cruzamento PMX (Goldberg, 1989).

Para situações, como a deste trabalho, onde foi necessário restringir o tamanho *S* do subconjunto que se deseja encontrar, no AG permutação considerou-se apenas as *S* primeiras posições para o cálculo da aptidão do cromossomo:

Subconjunto de tamanho = 2

4	1	2	3	6	5

No caso do AG binário aplica-se um fator de penalização na expressão da função de aptidão quando o número de atributos presentes no cromossomo não corresponde ao tamanho *S* do subconjunto de atributos desejado. A função adotada neste trabalho foi:

$$aptidao_{c/\,penalização} = b^{|d-S|} * aptidao_{s/\,penalização}$$

onde *b* = constante de penalização dada pelo usuário; *d* = número de bits acesos; *S* = tamanho do subconjunto de atributos desejado.

Não há um critério único para terminar a execução de um algoritmo genético; os critérios utilizados neste trabalho foram: determinação do número de gerações ou melhoras não significativas após *C* gerações, através da condição:

$$\left|X_N - X_{N-C}\right| < K\,X_N$$

onde X_N = média da população *N*; *C* = constante dada pelo usuário; *K* = percentual de adaptação com valores entre 0 e 1, dado pelo usuário.

4. Materiais e Métodos

Para avaliação dos resultados do trabalho foram realizados testes em imagens sintéticas e em uma imagem Quickbird de São José dos Campos com 0,6 m de resolução espacial. Para as duas imagens foram colhidas amostras das classes "Rua" e "Não-Rua". A imagem sintética, e as nove amostras retiradas para cada classe estão ilustradas nas **Figuras 1 e 2** respectivamente.

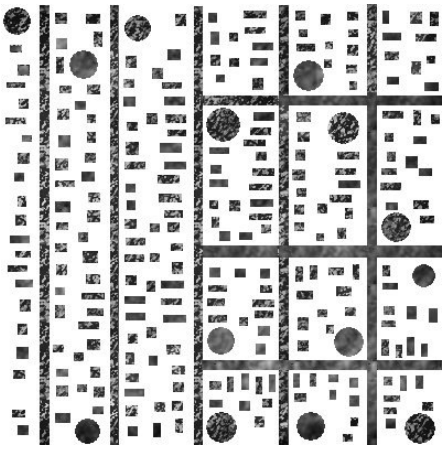
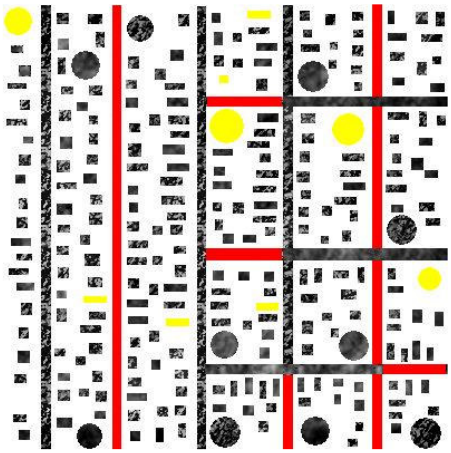


Figura 1 – Imagem sintética



Rua = * Não-Rua = *

Figura 2 - Imagem sintética com amostras

A imagem Quickbird de São José dos Campos juntamente com sua segmentação está na **Figura 5**. A segmentação da imagem de São José dos Campos foi feita manualmente dada a necessidade de se ter uma segmentação com formas bem definidas. O bom desempenho de um classificador de regiões é altamente dependente de uma boa segmentação. Na **Figura 6** são ilustradas as cinco amostras de cada classe utilizadas para o processamento da imagem Quickbird.

Para as classificações realizadas nos testes deste trabalho foi escolhido o classificador de distância euclidiana já disponível no sistema Texture. Para avaliação das classificações foi usado o coeficiente kappa que é obtido a partir da matriz de classificação, que fornece a

distribuição de pixels classificados correta e erroneamente (Cohen, 1960). A matriz de classificação é capaz de mostrar a exatidão do resultado de uma classificação comparando este resultado com a verdade de campo (imagens ou amostras).

6. Resultados e Discussão

No teste da imagem sintética, para a seleção de atributos foram usados os dois AGs citados anteriormente. De um conjunto de quarenta e seis atributos foi selecionado, pelos dois AGs, um subconjunto de quatro atributos. Para os dois métodos os parâmetros usados foram: população = 900 indivíduos; taxa de crossover = 0,6; taxa de mutação = 0,0333; critério de parada = melhoras não significativas com C = 5 e K = 0,01; constante de penalização = 0,7. Os gráficos da **Figura 3** mostram a evolução do indivíduo de maior aptidão e a evolução da média da aptidão da população em função das gerações.

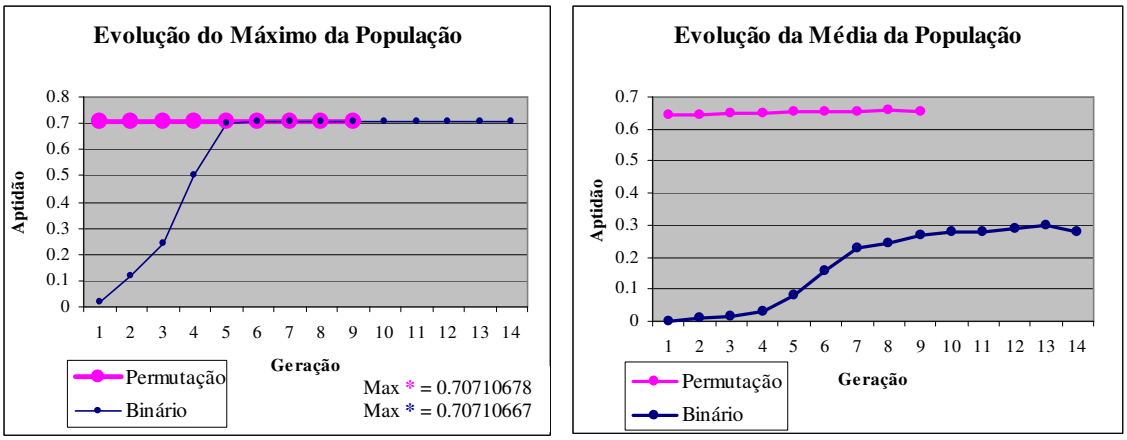


Figura 3 – Evolução da população para os AGs permutação e binário

É importante observar que no caso do AG permutação a população inicial, ainda que formada de modo aleatório, possui todos os indivíduos com excelente aptidão uma vez que seus cromossomos já satisfazem o critério do tamanho do subconjunto de atributos. No caso do AG binário, a sua população inicial, também gerada de forma aleatória, possui cromossomos com todos os tamanhos de subconjuntos de atributos, fazendo com que a aptidão média seja muito baixa. No entanto, dirigido pelas penalizações aos cromossomos inadequados, o AG binário evolui eficientemente para a solução desejada.

A partir dos gráficos da **Figura 3** é possível perceber que o método da permutação atinge resultados de forma mais rápida do que o método binário, como era de se esperar. O AG binário gastou quatorze gerações para encontrar a solução enquanto o AG permutação gastou nove gerações. É importante ressaltar que a solução encontrada pelo AG permutação é a solução ótima que foi encontrada pela busca exaustiva já existente no sistema. Para encontrar a solução através da busca exaustiva foram analisadas 135.751 combinações possíveis de atributos enquanto o AG permutação encontrou a solução analisando 8.100 (tamanho da população x número de gerações alcançado).

A classificação realizada (**Figura 4**) com os atributos encontrados na solução alcançou um coeficiente kappa = 0,9927. Para o cálculo deste coeficiente foi utilizada uma imagem verdade.

No teste da imagem Quickbird, para a seleção de três atributos foi utilizado o AG permutação com os seguintes parâmetros: população = 1.000 indivíduos; taxa de crossover = 0,6; taxa de mutação = 0,0333; critério de parada = melhoras não significativas com C = 5 e K

= 0,01. A classificação realizada com os atributos encontrados pelo AG permutação é mostrada na **Figura 7**. O coeficiente kappa encontrado para esta classificação foi de 0,9663. Para o cálculo deste coeficiente foram usadas amostras verdade ilustradas na **Figura 8**.

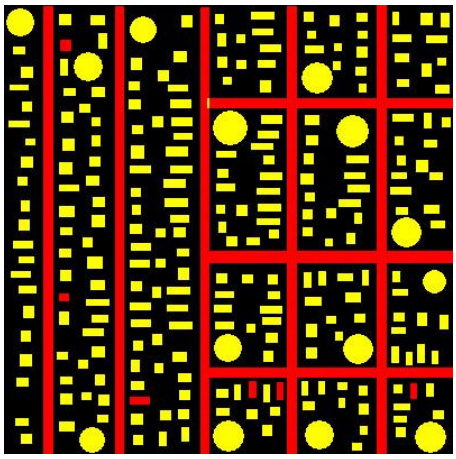


Figura 4 – Imagem sintética classificada

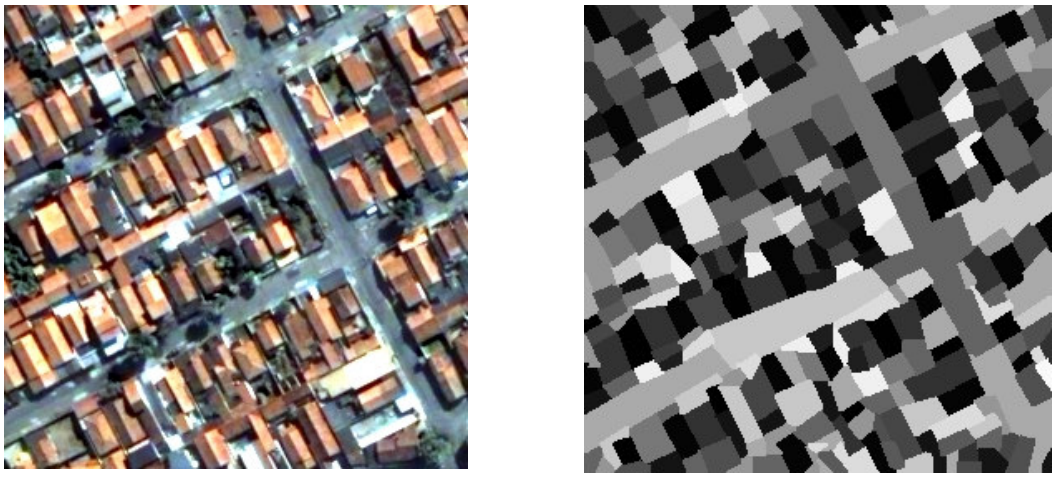


Figura 5 – Imagem Quickbird São José dos Campos e sua segmentação

Rua = * Não-Rua = *

Figura 6 - Imagem Quickbird com amostras

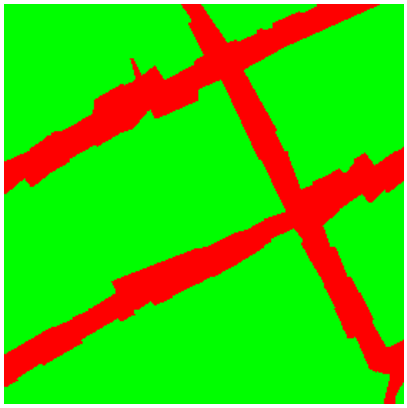


Figura 7 – Classificação com atributos AG Permutação

Para comparação com o resultado do AG permutação também foi feita uma classificação com três atributos que foram selecionados pela busca exaustiva. Esta classificação obteve um coeficiente Kappa de 0,8647 calculado também com as amostras verdade da **Figura 8**.

Na **Figura 9** é mostrada a classificação da imagem Quickbird usando todos os quarenta e seis atributos existentes no sistema. Observa-se nesta classificação uma confusão muito grande, ou seja, o número de regiões classificadas corretamente é menor do que na **Figura 7**.

Rua = * Não-Rua = *

Figura 8 – Amostras verdade

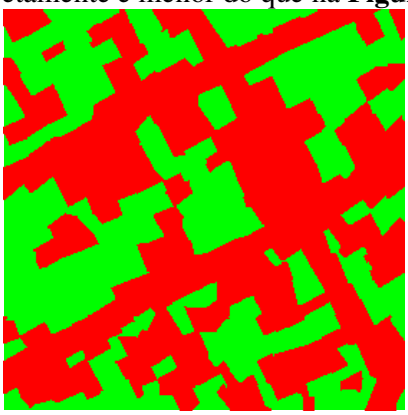


Figura 9- Imagem Quickbird classificada com 46 atributos

6. Conclusão

No processo de classificação, uma quantidade alta de atributos pode não aumentar a precisão do classificador; isso acontece devido à redundância das informações. Os resultados obtidos neste trabalho mostram a viabilidade em se utilizar métodos de seleção de atributos, objetivando a redução da dimensionalidade sem haver perda no poder discriminatório entre as classes. O AG permutação se mostrou um método mais eficiente e robusto do que o AG binário na solução deste problema. Também é importante dizer que obter uma boa classificação não depende apenas dos métodos de seleção de atributos, mas também do classificador utilizado. Um exemplo foi a classificação feita na imagem de São José dos Campos, onde foi selecionada pela busca exaustiva a combinação de atributos, entre todas as possíveis, a que oferecia a maior distância média JM entre as classes; porém a classificação obtida com os atributos selecionados pelo AG permutação obteve um coeficiente kappa maior.

Referências

Andrade, A. F.; Centeno, J. A. S.; Araki, H. Utilização de Parâmetros de Forma como Dado Auxiliar na Classificação de Imagens Ikonos através de Redes Neurais Artificiais. In: Simpósio Brasileiro de Geomática, 2002, Presidente Prudente, SP. **Anais...** Presidente Prudente: Unesp, 2002. Artigos, p. 342-349. Disponível em: <http://www.geomatica.ufpr.br/docentes/centeno/pessoal/download/2002/andreaf_unesp_A_091.pdf>. Acesso em: 15 dez. 2004.

Bhattacharyya, A.. On a measure of divergence between two statistical populations defined by their probability distributions. **Bulletin of the Calcutta Mathematics Society** 35, 99–110, 1943.

Boz, O.; Feature Subset Selection by Using Sorted Feature Relevance. In: International Conference on Machine Learning and Applications (ICMLA'02), 2002, Las Vegas, Nevada, USA. **Proceedings…** Las Vegas: CSREA Press, 2002. p. 147-153. (ISBN 1-892512-29-7) Disponível em: <

http://www.doc.ic.ac.uk/~xh1/Referece/feature-selection/Feature-Subset-Selection-by-Using-Sorted-Feature-Relevance.pdf > Acesso em: 7 jul. 2005.

Cohen, J. Coefficient of Agreement for Nominal Scales. **Educational and Psychological Measurement**. V.20 n.1 p. 37-43. 1960.

Donnay, J. P.; Barnsley, M. J.; Longley, P. A. Remote Sensing and Urban Analysis. In: Donnay, J. P.; Barnsley, M. J.; Longley, P. A. (ed) **Remote Sensing and Urban Analysis**. London: Taylor & Francis, 2001. cap 1, p.7-12p.

Duda, R. O.; Hart, P. E.; Stork, D. G. **Pattern Classification**. New York: Wiley, 2001. 654 p

Goldberg, D. E. **Genetic Algorithms in Search, Optimization, and Machine Learning**. Addison-Wesley, Nova York, 1989, 412p.

Lacerda, E.G.M.; Carvalho, A.C.P.L.F. Introdução aos Algoritmos Genéticos. In: Jornada de Atualização em Informática, XIX Congresso Nacional da Sociedade Brasileira de Computação, 1999b, Rio de Janeiro. **Anais...** Rio de Janeiro, 1999b. Artigos, p. 51-126.

Marques Filho, O. e Vieira Neto, H. Processamento Digital de Imagens, Brasport, 1999.

Oliveira, J. R. F. **O uso de algoritmos genéticos na decomposição morfológica de operadores invariantes em translação aplicados a imagens digitais**. 1998. 110 p. (INPE-10462-TDI/929). Tese (Doutorado em Computação Aplicada) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos. 1998.

Oliveira, J. A.; Dutra, L. V.; Rennó, C. D. **Classificação de Regiões Usando Atributos de Forma e Seleção de Atributos**. 2005. 99 p. Dissertação (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos - SP, 2005

Pappa, G. L. **Seleção de Atributos Usando Algoritmos Genéticos Multiobjetivos.** 2002. 85 p. Dissertação (Mestrado em Informática Aplicada) - Pontifícia Universidade Católica do Paraná. Curitiba, 2002a. Disponível: < http://www.ppgia.pucpr.br/ensino/defesas/Gisele_Lobo_2002.PDF >. Acesso em: 15 fev. 2005.

Pappa, G. L.; Freitas, A. A. and Kaestner, C. A. A. **A Multiobjective Genetic Algorithm for Attribute Selection**. In J. Garibaldi A Lofti and R. John, editors, Proc, 4th Int. Conf. on Recent Advances in Soft Computing (RASC-2002), pages 116-121. Nottingham Trent University, 2002b.

Rennó, C. D.; Freitas, C. C.; Frery, A. C. A system for region image classification based on textural measures. In: Jornada Latino-Americana de Sensoriamento Remoto por Radar: Técnicas de Processamento de Imagens, 2, 1998, Santos, SP. **Proceedings...** Noordwojk, ESA, 1998. Artigos, p. 159-164. Disponível em: <http://iris.sid.inpe.br:1908/rep-/sid.inpe.br/deise/1999/02.11.14.26> Acesso em: 16 abr. 2005.

Souza, I. M.; Pereira, M. N.; Fonseca, L. M. G.; Kurkdjian, M. L. N. O. Mapeamento do Uso do Solo Urbano através da Classificação por Regiões Baseada em Medidas Texturais. In: Simpósio Brasileiro de Sensoriamento Remoto, 11., 2003, Belo Horizonte. **Anais...** São José dos Campos: INPE, 2003. Artigos, p. 1967-1968. Disponível em: < http://iris.sid.inpe.br:1908/col/ltid.inpe.br/sbsr/2002/11.14.15.20/doc/14_187.PDF>. Acesso em: 26 abr 2005.

Vafaie, H. and De Jong, K. (1992). Genetic Algorithms as a Tool for Feature Selection in Machine Learning. In International Conference on Tools with Artificial Intelligence, 1992, Arlington. **Proceedings…** Arlington, 1992. P. 200-204. Disponível em: <http: //cs.gmu.edu/~eclab/papers/TAI92.pdf> Acesso em: 15 dec. 2004

Vafaie, H. & De Jong, K. (1993) Improving the Performance of a Rule Induction System Using Genetic Algorithms. In: **Machine Learning: A Multistratey Approach**, 1993. Edited by R.S. Michalski and G. Tecuci, San Mateo, CA: Morgan Kaufmann.