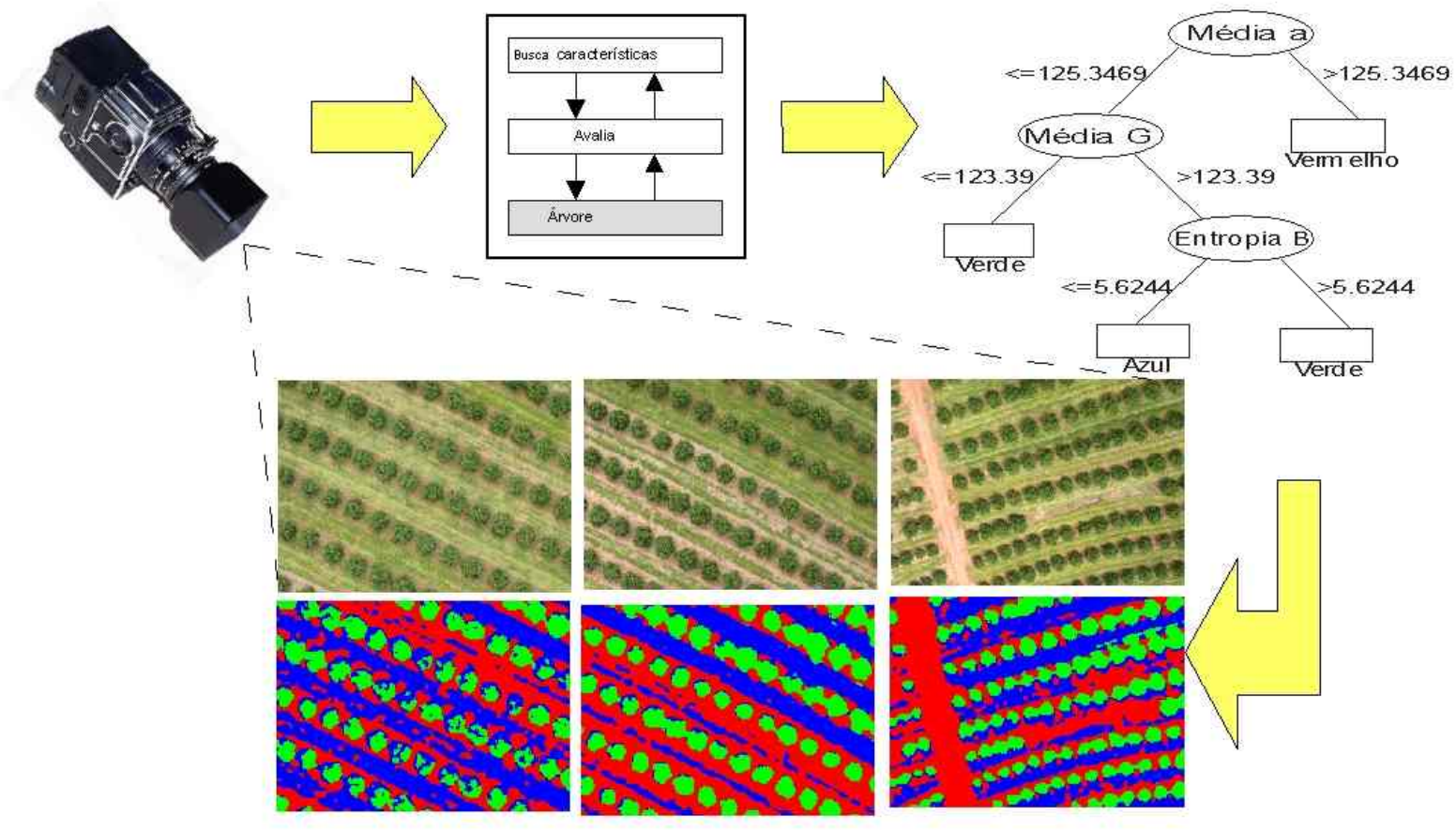
Documentos

ISSN 1518-7179
Novembro, 2007
33

Seleção de Características Aplicada ao Processamento de Imagens Digitais



Embrapa

ISSN 1518-7179
Novembro, 2007

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Instrumentação Agropecuária
Ministério da Agricultura e do Abastecimento

Documentos 33

Seleção de Características Aplicada ao Processamento de Imagens Digitais

Ednaldo José Ferreira
Lúcio André de Castro Jorge

Embrapa Instrumentação Agropecuária
São Carlos, SP
2007

Exemplares desta publicação podem ser adquiridos na:

Embrapa Instrumentação Agropecuária
Rua XV de Novembro, 1452
Caixa Postal 741
CEP 13560-970 - São Carlos-SP
Fone: (16) 3374 2477
Fax: (16) 3372 5958
www.cnpdia.embrapa.br
E-mail: sac@cnpdia.embrapa.br

Supervisor editorial: Dr. Victor Bertucci Neto
Normalização bibliográfica: Valéria de Fátima Cardoso
Tratamento de ilustrações: Valentim Monzane
Imagem da capa: Lúcio André de Castro Jorge
Editoração eletrônica: Valentim Monzane

1ª edição
1ª impressão (2007): tiragem 300

CIP-Brasil. Catalogação-na-publicação.
Embrapa Instrumentação Agropecuária

F383s Ferreira, Ednaldo José
Seleção de Características Aplicada ao Processamento de Imagens Digitais. / Ednaldo José Ferreira, Lúcio André de Castro Jorge. - São Carlos: Embrapa Instrumentação Agropecuária, 2007.
21 p. - (Embrapa Instrumentação Agropecuária. Documentos, ISSN 1518-7179; 33).

1. Inteligência artificial - Seleção de características. 2. Imagens Modelos de cores. 3. Imagens Segmentação. 4. Inteligência artificial - Wrapper. I. Jorge, Lúcio André de Castro. II. Título. III. Série.

CDD 21 ED 006.31
006.4

Autores

Ednaldo José Ferreira
Ciência da Computação, MSc., Embrapa Instrumentação Agropecuária,C.P.741,
CEP 13560-970, São Carlos (SP)
ednaldo@cnpdia.embrapa.br

Lúcio André de Castro Jorge
Eng. Eletrecista, MSc., Embrapa Instrumentação Agropecuária,C.P.741,
CEP 13560-970, São Carlos (SP)
lucio@cnpdia.embrapa.br

Apresentação

O uso de técnicas de mineração de dados e aprendizado de máquina no processo de descoberta de conhecimento de base de dados de diversos domínios tem proporcionado modelos de predição mais precisos, exatos e, dependendo da concepção, auto-explicativos. O processo de seleção de características relevantes é umas das tarefas de pré-processamento que contribui significativamente para a redução da dimensão dos dados, maior acurácia e melhor compreensão dos modelos gerados, aspectos que tornam a seleção de características uma ferramenta importante na construção dos modelos preditivos.

No processo de segmentação de imagens, o número de características extraídas de uma imagem digital não assegura a eficiência dos modelos adotados para predição. A característica (atributo) "componente de cor", por exemplo, pode estar altamente correlacionada a outras, principalmente se diversos modelos de cores forem considerados na segmentação. Por outro lado, características podem ter alto custo computacional de extração e apresentarem-se irrelevantes ao modelo de predição adotado.

Estudos indicam que a presença de muitas características irrelevantes deteriora o desempenho dos algoritmos na construção dos modelos de predição. Por isso, métodos adicionais de seleção de características representam uma alternativa importante para aumentar a acurácia dos modelos gerados, reduzir custos de armazenamento e viabilizar a análise de características. Nesse contexto, o objetivo desta série é apresentar os principais conceitos e abordagens de seleção de características, assim como sua aplicação em segmentação de imagens aéreas de fazendas de citrus visando a identificação e gerenciamento da área cultivada.

Álvaro Macedo da Silva
Chefe Geral

Sumário

Introdução **9**

Seleção de Subconjunto de Características **9**

Seleção de Características como Busca Heurística **11**

Ponto de partida 12

Organização da busca 12

Estratégia de Avaliação 12

Critério de parada 12

Abordagens para Seleção de Características **13**

Embutida 13

Filtro 13

Wrapper 14

Seleção de Características no processamento de imagens **15**

Referências **21**

Seleção de Características Aplicada ao Processamento de Imagens Digitais

Ednaldo José Ferreira
Lúcio André de Castro Jorge

Introdução

A evolução da tecnológica tem propiciado a coleta e armazenagem de grandes quantidades de dados em diversas áreas. Estima-se que a quantidade de dados mundialmente gerados duplique a cada dois anos. Esses dados podem ser usados para descrever e caracterizar conceitos assim como, por exemplo, a cor, a forma e textura podem descrever uma classe de objetos em imagens digitais. Sobre esses dados, algoritmos de aprendizado do paradigma supervisionado podem induzir uma hipótese (modelo) para predizer os rótulos de novos exemplos. O sucesso para construção de um modelo preditor adequado, seja ele um classificador ou um modelo para regressão, depende da relevância das características do conjunto de dados. Características irrelevantes interferem na construção desses modelos e podem acarretar baixa acurácia de predição.

A seleção de características relevantes é um dos temas pesquisados em Aprendizado de Máquina (AM). Embora a maioria dos algoritmos de AM busque, intrinsecamente, selecionar características ou atribuir-lhes graus de importância, estudos indicam que a presença de muitos atributos irrelevantes deteriora o desempenho desses algoritmos na construção das hipóteses. Por isso, métodos adicionais de seleção de características representam uma alternativa para aumentar a acurácia das hipóteses geradas, reduzir custos de armazenamento e viabilizar a interpretação de atributos e dos modelos gerados. Nesse contexto, o objetivo desta série é apresentar os principais conceitos e abordagens de seleção de características; outrossim, apresentar as potencialidades de aplicação em processamento de imagens digitais.

Seleção de Subconjunto de Características

A qualidade de uma hipótese induzida por um algoritmo de aprendizado depende da relevância das características consideradas no conjunto de exemplos de treinamento T. Em problemas de classificação, por exemplo, o conjunto de exemplos T é composto por m instâncias (x, y), onde o vetor x é um elemento do produto cartesiano F_1 x F_2 x ... x F_I; onde y é a classe que x pertence; F_i é o domínio da i-ésima característica e I é o número de características utilizadas para descrever o exemplo em questão. A relevância das características é fundamental para o algoritmo de AM no aprendizado de conceitos. Quanto maior a quantidade de características irrelevantes, maior a necessidade de exemplos de treinamento para alcançar uma dada acurácia (LANGLEY e IBA,1993).

Em síntese, a tarefa de aprendizado de conceitos pode ser divida em (1) decidir que características utilizar na descrição do conceito e (2) decidir como combiná-las (BARANAUSKAS, 2001). Os métodos de seleção de subconjunto de características (FSS - *Feature Subset Selection*) têm, portanto, a missão de encontrar uma combinação adequada de características relevantes para o aprendizado de um conceito. A relevância de características é um conceito de definição complexa. Características individualmente irrelevantes podem ser relevantes quando combinadas. A relevância depende também das peculiaridades do modelo de aprendizado (algoritmo indutor) adotado. Assim, um subconjunto relevante para um determinado modelo, pode não ser similar para outro modelo de aprendizado distinto. Diversas definições para relevância de características podem ser encontradas em Blum e Langley (BLUM e LANGLEY, 1997).

Um processo de FSS é caracterizado pela redução da dimensão do vetor de características ou, em teoria de conjuntos, a redução do conjunto *U* das características. A partir de *U*, um subconjunto de características *R* é selecionado usando algum critério, tal que (*U-R)?Ø*. Em outras palavras, o conjunto original *U* é analisado e processado para produzir um subconjunto *R* com um número menor de atributos. Várias razões justificam e motivam a aplicação de FSS (LEE et al., 1999). A primeira razão é que alguns algoritmos de AM computacionalmente viáveis não trabalham adequadamente na presença de um grande número de características, principalmente se houver características irrelevantes. Isso significa que a FSS é um instrumento importante para aumentar a acurácia dos classificadores gerados por esses algoritmos. Outra razão é que a FSS pode auxiliar na compreensão dos dados pois, pode indicar características importantes para o processo de predição. A terceira razão é que FSS pode ajudar a reduzir os custos envolvidos na coleta de dados de alguns domínios. Finalmente, a redução de características por FSS pode reduzir significativamente o custo de processamento de grandes quantidades de dados.

O processo de FSS pode ser descrito como uma busca em um espaço de estados, onde cada estado representa um subconjunto de características que é mensurável por meio de critérios pré-estabelecidos para prover uma estimativa da relevância daquele subconjunto. Operadores de adição ou remoção de características são comuns e possibilitam a travessia ou caminhamento no espaço de estados. Na Figura 1 é mostrado um grafo que representa o espaço de busca para quatro características, onde cada nó corresponde a um estado.

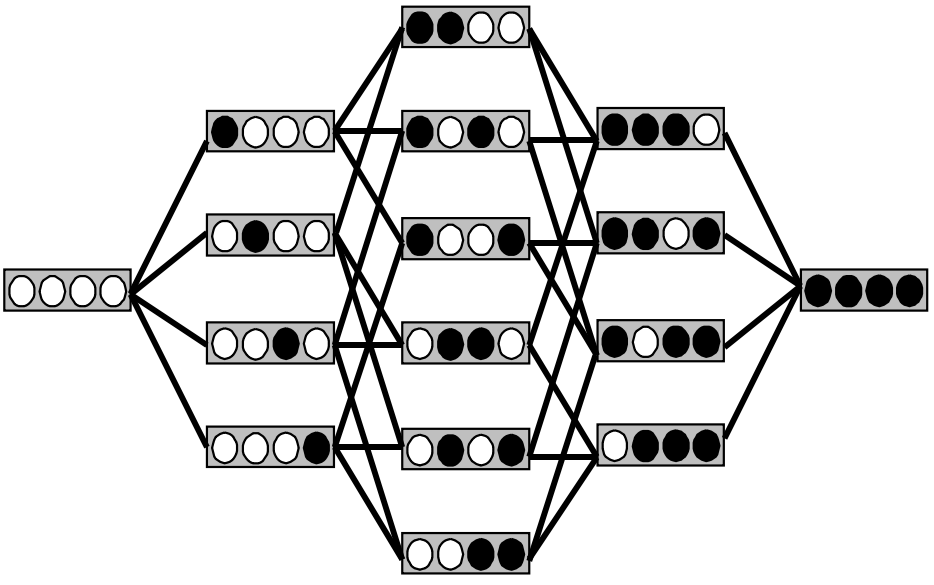


Fig. 1 - Espaço de busca (estados) (BLUM e LANGLEY, 1997).

Os círculos em preto representam a presença das características e os círculos em branco a ausência delas. Cada estado no espaço dos subconjuntos de características especifica quais são os atributos candidatos à indução de um classificador. Os estados no espaço de busca são estruturados (ordenados) por número de características. Em cada nível, caminhando da esquerda para direita, uma característica é acrescentada. Os extremos representam a ausência e a presença total das características do problema em questão.

Percorrer todo o espaço de busca para encontrar o melhor subconjunto de características é uma tarefa computacionalmente inviável na grande maioria dos casos, pois a complexidade computacional é da ordem de 2^I, onde I é o número de características do conjunto U. Para exemplificar, suponha um conjunto de dados médicos descrito por 50 características (I = 50) representando os sintomas dos pacientes e os resultados dos exames que viabilizam o diagnóstico de um grupo de doenças hematológicas. Se a estimativa de precisão de um estado do espaço levasse um milésimo de segundo para ser computada e se a avaliação de todos os possíveis estados fosse requerida, o melhor resultado levaria cerca de 35.702 anos para ser emitido. Portanto, a busca exaustiva garante a melhor solução, mas é impraticável na maioria dos casos.

Para tornar a FSS viável, diversas soluções para seleção das características relevantes são encontradas na literatura. Para geração de novos candidatos há três abordagens:

- completa - é a busca por todos os estados no espaço de busca. Conforme descrito anteriormente é impraticável quando o número de características é relativamente grande. Sua complexidade é $O(2^I)$;
- heurística. - em cada iteração, todas as características ainda não selecionadas (ou selecionadas) são consideradas e avaliadas para seleção (ou remoção). A utilização de heurísticas pode diminuir drasticamente a complexidade da busca. Sua complexidade é $O(I^2)$;
- randômica - é uma abordagem relativamente nova comparada as outras duas categorias. Em geral, os subconjuntos candidatos são aleatoriamente amostrados do espaço de busca. Embora sua complexidade seja $O(2^I)$, os métodos dessa categoria efetuam, tipicamente, uma busca em um número limitado de subconjuntos de características. Algoritmos genéticos aplicados a FSS são excelentes alternativas desta categoria.

Um paradigma conveniente para visualizar e elucidar as abordagens de seleção de características é a busca heurística, apresentada a seguir.

Seleção de Características como Busca Heurística

A visualização das abordagens de seleção de características como uma busca heurística permite que qualquer método de FSS seja caracterizado em relação a quatro pontos básicos:

- ponto de partida;
- organização da busca;
- estratégia da busca;
- critério de parada.

---

[1]Esta estratégia é denominada de busca exaustiva.

[2]Métodos que selecionam sempre o estado com a melhor estimativa a cada iteração.

Ponto de partida

O ponto de partida determina qual é o estado no espaço onde a busca deve iniciar. A determinação do ponto de partida influencia na direção da busca e nos operadores que serão utilizados para geração dos estados sucessores. Pode-se observar no grafo da Figura 1 a existência de um ordenamento lógico entre os estados, pois o estado adjacente à direita tem uma característica adicional e o estado adjacente à esquerda tem uma característica subtraída. Se o ponto inicial é determinado pelo estado sem nenhuma característica (primeiro estado à esquerda) e o operador adiciona características gradativamente, então a abordagem é conhecida como seleção *forward*. Se o ponto inicial é o estado com todas as características (último estado à direita) e o operador remove características gradativamente, então abordagem é conhecida como seleção *backward*. Além disso, pode ser empregada uma variação de ambas, escolhendo-se como ponto de partida qualquer estado do espaço e movendo-se a partir desse ponto seleção *outward*.

Organização da busca

Conforme descrito anteriormente, a busca exaustiva por todos os estados no espaço de busca é impraticável. Uma abordagem mais prática e viável é a utilização de métodos gulosos para travessia do espaço de busca: em cada estado do espaço consideram-se alterações locais sobre o atual subconjunto de características, seleciona-se uma nova característica e realiza-se uma nova iteração. A abordagem *hill-climbing*, por exemplo, conhecida por seleção ou eliminação *stepwise*, considera tanto a adição quanto a remoção de características em cada ponto (estado) de decisão, além de permitir que um estado anterior possa ser reconsiderado durante a busca.

Estratégia de Avaliação

Um dos pontos fundamentais na busca é a estratégia utilizada para avaliar o subconjunto de características. Uma métrica geralmente empregada é a capacidade que uma característica tem para discriminar as classes em um conjunto de treinamento. Diversos algoritmos empregam um critério baseado na Teoria da Informação, enquanto outros medem diretamente a acurácia do classificador sobre o conjunto de treinamento ou sobre um conjunto de teste.

A estratégia de avaliação está relacionada à maneira com que o algoritmo de FSS lida com o algoritmo básico de AM. A independência de avaliação com respeito ao algoritmo de AM é um característica fundamental da abordagem *filtro*. A dependência do algoritmo de AM caracteriza a abordagem conhecida como *wrapper*.

Critério de parada

A dimensão do espaço de estados pode tornar o processo de seleção de características computacionalmente custoso, mesmo na busca heurística. Assim, algum critério para parar a busca deve ser definido. O critério de parada pré-estabelece as condições em que a busca deve cessar. Alguns critérios utilizados são:

- parar de adicionar ou remover características quando nenhuma das alternativas melhora a acurácia estimada do classificador;
- parar quando um dos extremos do espaço de busca for alcançado e selecionar a melhor solução;
- parar quando um número pré-definido de características é alcançado;
- parar após um número pré-definido de iterações.

Abordagens para Seleção de Características

Segundo Blum e Langley (1997), os métodos de FSS podem ser agrupados em três classes/abordagens: embutida, filtros e *wrapper*. Cada uma dessas abordagens é discutida nas seções subseqüentes.

Embutida

Na indução de hipóteses, alguns algoritmos de aprendizado podem realizar uma seleção dinâmica de características. Nesses casos, a FSS é um componente intrínseco do algoritmo de aprendizado.

Os métodos de particionamento recursivo, tais como as árvores de decisão, efetuam uma busca gulosa através do espaço de estados/características. A cada passo, uma função de avaliação determina qual característica tem maior capacidade de discriminação das classes. Com base na característica escolhida, o conjunto de treinamento é particionado e o processo repetido, estendendo a árvore até que nenhuma discriminação adicional seja possível. Esse tipo de abordagem é usada pelo algoritmo C4.5 (QUINLAM, 1993). Na Figura 2 é apresentada uma ilustração da abordagem de FSS embutida para um algoritmo de construção de uma árvore de decisão. Observa-se que a árvore de decisão final (hipótese) utiliza apenas uma parcela das características (representadas como f_2, f_6 e f_7) do conjunto total.

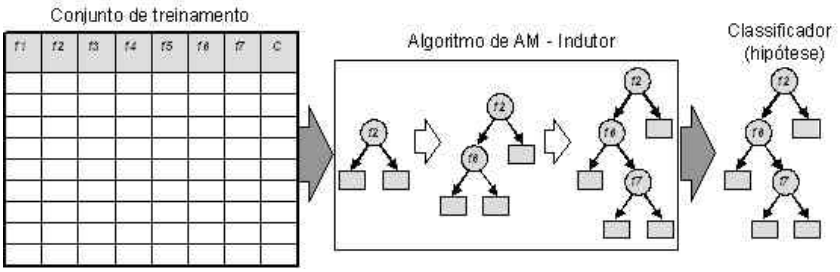


Fig. 2 - Abordagem embutida.

Filtro

Essa abordagem de seleção de características é caracterizada pela independência de seleção em relação ao algoritmo de AM utilizado para gerar o modelo final. Os filtros atuam na fase de pré-processamento dos dados e fazem uma seleção de características, recebendo como entrada o conjunto de treinamento, com todas as características, avaliando as propriedades intrínsecas dos dados e produzindo um subconjunto de maior relevância segundo algum critério estatístico. Na Figura 3 é ilustrada a abordagem filtro.

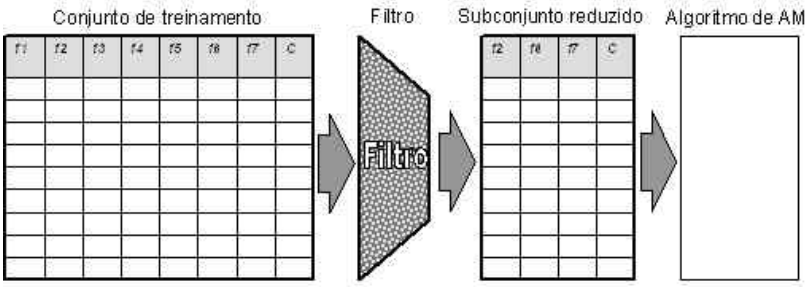


Fig. 3 - Abordagem Filtro.

Um dos esquemas mais simples de filtragem é a avaliação de cada característica individualmente, baseada na sua correlação com a classe (*C*), escolhendo as *r* características com melhor valor.

A aplicação de algoritmos de filtragem oferece vantagens, tais como: eficiência para lidar com grande quantidade de dados, baixo custo de processamento computacional e independência do algoritmo de AM. A principal desvantagem dessa abordagem é que os filtros ignoram totalmente os efeitos do subconjunto de características no desempenho do algoritmo de AM ao qual os dados reduzidos serão submetidos (BARANAUSKAS, 2001). Por isso, Kohavi e John (1997) propõem uma abordagem, denominada *wrapper*, que considera o próprio algoritmo de AM no processo de seleção.

Wrapper

Os algoritmos baseados nessa abordagem geram um subconjunto de características como candidato, executam o algoritmo de AM com os dados de treinamento desse subconjunto e medem a acurácia do modelo gerado. Uma medida de acurácia do modelo é utilizada para avaliar o subconjunto de características em questão. Esse processo é repetido a cada novo subconjunto gerado até que algum critério de parada seja satisfeito. Todo processo envolvido na abordagem *wrapper* é apresentado na Figura 4.

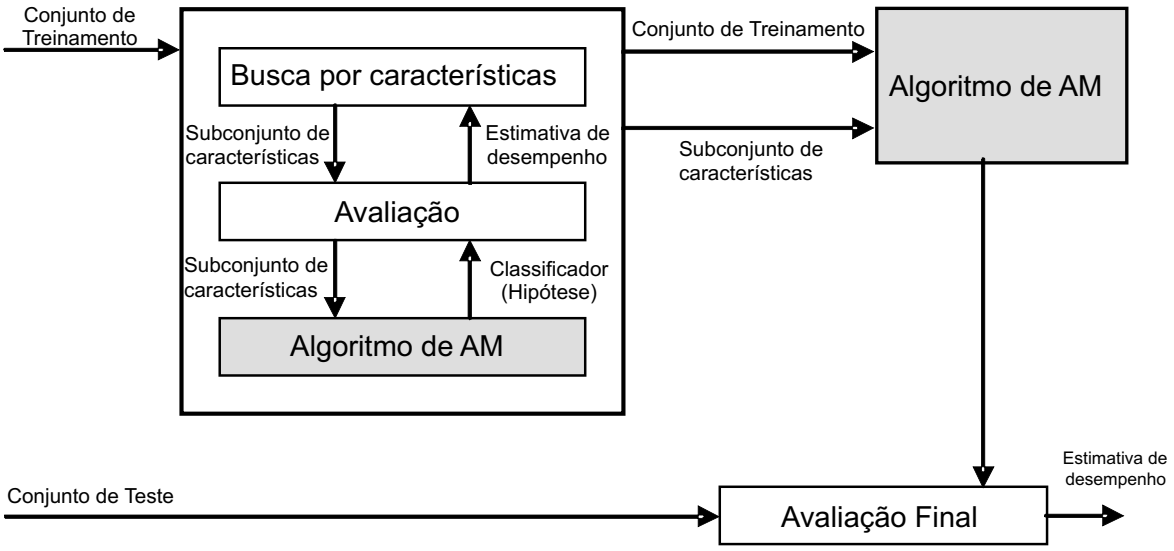


Fig. 4 - Abordagem Wrapper (KOHAVI e JOHN, 1997).

A principal desvantagem dessa abordagem é o custo computacional. Para cada subconjunto de características, um modelo deve ser gerado e sua acurácia estimada. Esse fator encarece computacionalmente todo o processo de busca pelo subconjunto adequado. Entretanto, a principal vantagem é que o subconjunto de características selecionado é otimizado para o modelo final. Por essa razão, *wrapper* geralmente supera a abordagem filtro.

Quando o espaço de busca é demasiadamente grande, torna-se impraticável o uso de técnicas de busca tradicionais em *wrapper* como, por exemplo, *hill-climbing* com algoritmos de AM de RNAs. Assim, algoritmos genéticos são, geralmente, mais eficazes para uma rápida busca global em grandes espaços de busca, além de oferecerem um atrativo para uma otimização multi-objetivo (YANG e HONAVAR, 1998).

Seleção de Características no processamento de imagens

A seleção de características em uma imagem tem grande importância para sistemas de visão computacional. Uma grande quantidade de informações presentes em uma imagem, como por exemplo, a cor, a textura e a forma, podem ser ou não relevantes para o processo de segmentação, classificação ou análise. Um exemplo é a cor, que é uma das características mais óbvias e importantes da percepção, sendo um atributo de sensação visual devido à interação de três componentes: fontes de luz, objeto e o sistema de visão.

Trabalhar com imagens coloridas, requer escolher o sistema de cores mais adequado ou uma combinação de sistemas de cores. O uso de diversos modelos de cores é uma tarefa complexa para os algoritmos de visão computacional, pois o espaço de cor pode ser interpretado e modelado de diferentes formas. Nestes casos, combinar os espaços de cores ou os canais de cores ou mesmo, achar a melhor combinação desses espaços para uma determinada aplicação é um grande desafio. É possível que diversos espaços de cores sejam bons candidatos no processo de segmentação ou classificação, pois podem possuir propriedades similares, como por exemplo, os canais V e o G, do modelo RGB e HSV (JORGE et al., 2007), trazem informações sobre intensidade da cor verde.

Sendo assim, selecionar os atributos, seja de cor ou outros, não é um processo trivial e é ai que as técnicas de seleção de características advindas do aprendizado de máquina podem ajudar sensivelmente.

Na Figura 5, podem ser observadas as etapas da seleção de características para o processamento de imagens. Primeiramente, se adquire as imagens do campo de aplicação, em seguida se aplicam as diferentes técnicas existentes para extrair atributos ou características importantes. Estas características podem ser: a cor, a textura, a forma, um histograma, as componentes principais, dentre outras, dependendo do objetivo do processamento, por exemplo, a segmentação, a classificação, a análise, a busca e recuperação de informações num banco de imagens, etc. Uma vez extraídas as diferentes características, a mineração de dados permite a seleção das características mais significativas para a aplicação.

Apenas para exemplificar, continuando na escolha das componentes de cores, a grande dificuldade é selecionar automaticamente um subconjunto ótimo de componentes de cores e suas características mais representativas, produzindo o melhor resultado para determinada aplicação, o que é conseguido através do balanço adequado entre cores invariantes (repetibilidade) e variantes (poder discriminatório).

Baseando-se na noção de separabilidade de classes, foi aplicado o método *Wrapper* para selecionar um subconjunto de componentes de cores visando conseguir a melhor segmentação da imagem segundo os modelos de cores selecionados (JORGE et al., 2007). As etapas aplicadas para seleção das características, ou seja, das componentes de cor estão apresentadas na Figura 6.

Para se obter a seleção de das componentes de cor, foram definidos inicialmente dois conjuntos de imagens, o primeiro com imagens artificiais obtidas em condições controladas, e o segundo com imagens de uma aplicação real.

O primeiro conjunto contou com imagens de tamanho 192x144 pixels do objeto 25 da biblioteca de imagens da universidade de Amsterdã (Amsterdam Library of Objects Images - ALOI) (GEVERS, 2001). Essas imagens foram capturadas em diversas condições de iluminação, rotação e temperatura de cor. Alguns exemplos podem ser observados na Figura 7:

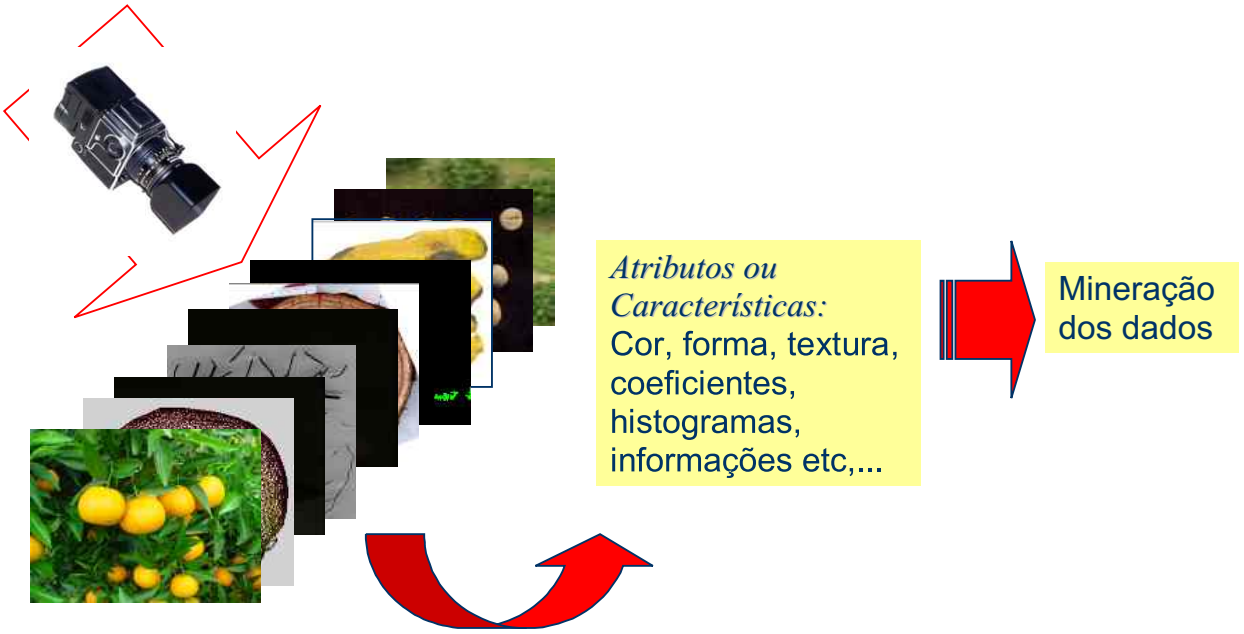


Fig. 5 - Seleção de características no processamento de imagens

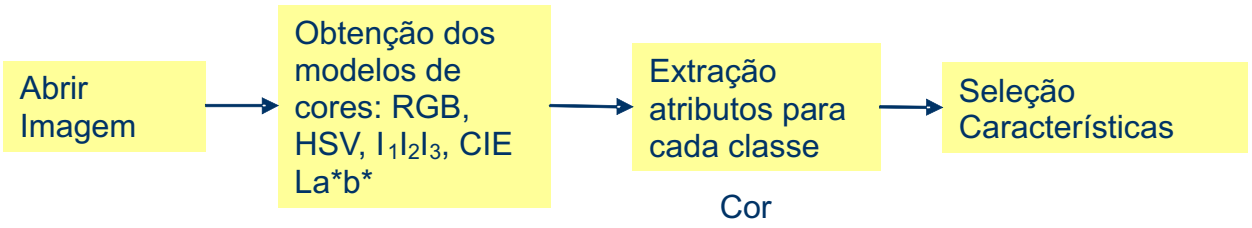


Fig. 6 - Um exemplo de seleção de características (componentes de cor) no processamento de imagens

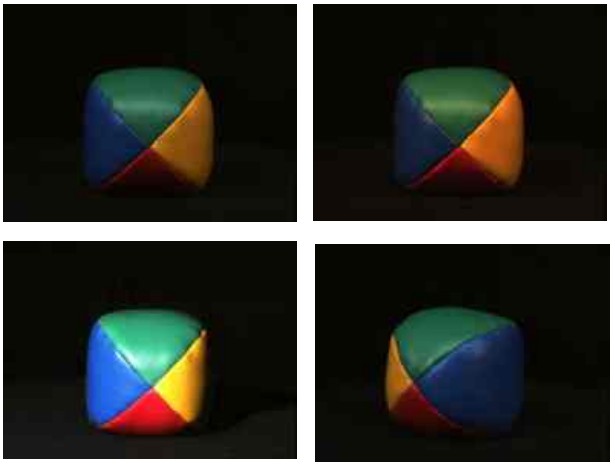


Fig. 7 - Objeto 25 da Biblioteca ALOI, apresentando mudanças de temperatura de cor, iluminação e rotação.

O segundo conjunto conta com imagens aéreas de plantações de citrus 372x248 pixels que foram adquiridas a 100 metros de altura, a diferentes posições e condições de iluminação. A Figura 8 mostra as imagens utilizadas na aplicação real.

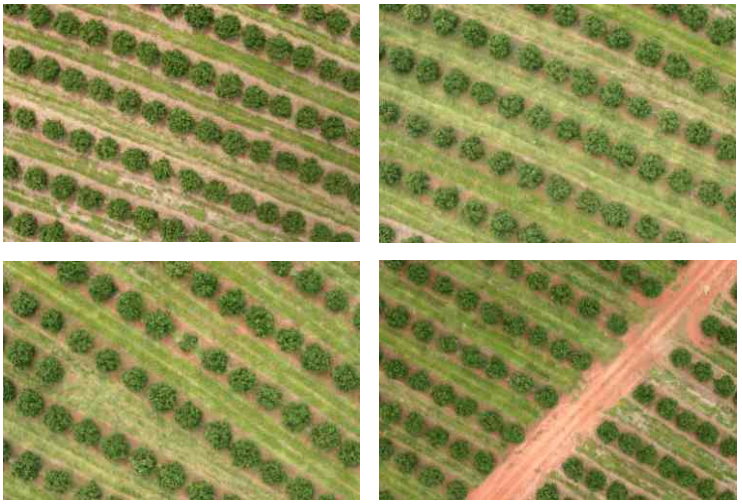


Fig. 8 - Imagens aéreas de fazendas de citrus.

A qualidade da classificação e segmentação das imagens, induzida por um algoritmo de aprendizado de máquina (AM) depende da relevância das características consideradas no conjunto de padrões de treinamento, chamado de vetor de características (VC). A relevância das características é fundamental para o algoritmo de AM no aprendizado de conceitos. Quanto maior a quantidade de características irrelevantes, maior a necessidade de exemplos de treinamento para alcançar uma dada acurácia (LANGLEY e IBA, 1993).

Os métodos de seleção de características têm a missão de encontrar uma combinação adequada de características relevantes para o aprendizado de um conceito. Várias razões justificam e motivam a aplicação da seleção de características. A primeira razão é que alguns algoritmos de AM computacionalmente viáveis não trabalham adequadamente na presença de um grande número de características, principalmente se houver características irrelevantes. Isso significa que a seleção pode aumentar a acurácia dos classificadores gerados por esses algoritmos. Outra razão é que a seleção pode auxiliar na compreensão dos dados. A terceira razão é que a seleção pode ajudar a reduzir os custos envolvidos na coleta de dados de alguns domínios.

No exemplo, foram utilizados 4 modelos de cores, RGB, HSV, CIE L*a*b* e $I_1I_2I_3$ (JORGE et al., 2007). Os modelos HSV, CIE L*a*b* e $I_1I_2I_3$, foram obtidos a partir de transformações do modelo RGB. Esses modelos, tipicamente encontrados na literatura, contêm algumas propriedades variantes e invariantes no que diz respeito às condições fotométricas. As componentes de cor RGB, CIE L*, e SV são sensíveis às sombras, iluminação e brilho, ou seja, possuem um grande poder discriminatório. Já as componentes CIE a*b* são invariantes às sombras e intensidade da iluminação, possuindo grande capacidade de repetibilidade (GEVERS, 2001). Para o método proposto deve haver um balanço entre o poder discriminatório e a repetibilidade.

O primeiro experimento foi conduzido com 4 imagens do objeto 25 da biblioteca de imagens da Universidade do Amsterdã (ALOI). Este objeto é uma bola com as cores vermelho, azul, verde, amarelo e um fundo escuro, em diferentes condições de iluminação, rotação e temperatura de cor. Para a execução do método *Wrapper*, foi necessário capturar amostras dos padrões de cada classe presente nas imagens. As classes foram definidas por regiões r de cores diferentes, ou seja, vermelho, azul, verde, amarelo e o fundo. Foram capturadas 5 amostras de tamanho 3x3 de cada região, procurando contemplar as diversas condições de iluminação e brilho na seleção.

Para cada região *r* da imagem, foi tomada uma amostra do padrão de tamanho IxJ. Para cada uma das componentes de cor calculou-se a média, variância e entropia das amostras *r*(*i, j*), para *i = 1…I, j = 1…J*, como mostrado nas equações (1), (2) e (3). A entropia foi determinada pelo histograma *h*(*k*) onde *v*(h(*k*)) são as ocorrências dos valores e *k* é o nível de cinza em cada região *r*(*i,j*); considerando cada componente uma nova imagem em tons de cinza.

$$Média \quad \frac{1}{MN} \sum_{i\ 1}^{M} \sum_{j\ 1}^{N} r(i,j) \tag{1}$$

$$Variância \quad \frac{1}{MN} \sum_{i\ 1}^{M} \sum_{j\ 1}^{N} (r(i,j) \quad Média_{ij})^2 \tag{2}$$

$$Entropia \quad \sum_{i\ 0}^{255} h(k) * v(h(k)) \tag{3}$$

Depois de extraídos os valores da média, variância e entropia, o vetor de características VC foi arranjado conforme na seqüência:

VC = [MédiaR, MédiaG, MédiaB, VariânciaR, VariânciaG, VariânciaB, EntropiaR, EntropiaG, EntropiaB, MédiaL, Médiaa, Médiab, VariânciaL, Variânciaa, Variânciab, EntropiaL, Entropiaa, Entropiab, MédiaI1, MédiaI2, MédiaI3, VariânciaI1, VariânciaI2, VariânciaI3, EntropiaI1, EntropiaI2, EntropiaI3, MédiaH, MédiaS, MédiaV, VariânciaH, VariânciaS, VariânciaV, EntropiaH, EntropiaS, EntropiaV].

Após a criação do vetor de características VC, o método *Wrapper* com a busca exaustiva foi aplicado para selecionar o melhor subconjunto de características. Para avaliar o desempenho do subconjunto selecionado, as imagens de teste foram segmentadas utilizando as componentes selecionadas pelo *Wrapper* e a árvore gerada pelo algoritmo C4.5 (QUINLAN, 1993).

O segundo experimento foi realizado com um conjunto de imagens aéreas de fazendas de citrus, com o objetivo de identificar e gerenciar a qualidade da área cultivada. A segmentação dessas imagens poderá ser usada para identificar diferenças relativas ao vigor da cultura, às pragas, às doenças e ao nível de desenvolvimento da planta. Esse tipo de aplicação é particularmente interessante porque não é possível controlar a luz, as sombras e os reflexos do sol das imagens obtidas. Como na primeira experiência, os padrões de teste das amostras da região foram extraídos, porém usando seleções de tamanho diferentes. As classes selecionadas como padrão no teste, são: árvore de citrus, solo descoberto e ervas daninhas ou invasoras.

Como no primeiro experimento, para cada região r da imagem, é tomada uma amostra do padrão de cor de tamanho (I,J). Para cada uma destas componentes de cor calcula-se a média, variância e entropia das amostras *r(i, j)*, para *i = 1…I, j = 1…J*, como mostrado nas equações (1), (2) e (3). Novamente, o vetor *VC* foi arranjado para a tarefa de seleção de características, conforme o experimento anterior.

O método *Wrapper* foi testado nos conjuntos de imagens apresentados, em diversas condições de iluminação, rotação e temperatura de cor.

Iniciando pelo *Wrapper* com o algoritmo C4.5 por busca exaustiva, as imagens do objeto ALOI 25 foram processadas a partir das 12 médias das respectivas componentes de

cores. Os objetos apresentados na Figura 7, apresentam variação da temperatura de cor, reflexos pela iluminação e mudança do ponto de observação devido à rotação do objeto.

As componentes selecionadas foram I_3, H, L e V dos modelos de cores. Em uma segunda etapa, adicionou-se a variância e a entropia correspondente às componentes selecionadas I_3, H, L e V. Em todo o processo foi utilizada a validação cruzada em 10 desdobramentos (10-fold-cross-validation) nas amostras de treinamento, resultando em uma taxa de acertos de 99.62%. A árvore de decisão gerada é apresentada na Figura 9 e os resultados da segmentação podem ser observados na Figura 10.

Apesar das novas características adicionadas (variâncias e entropias), foi obtido o mesmo subconjunto das componentes de cores I_3, H, L e V, e conseqüentemente a mesma árvore da decisão. Conforme esperado, devido às características das imagens, as variâncias e as entropias não contribuíram para uma melhor classificação.

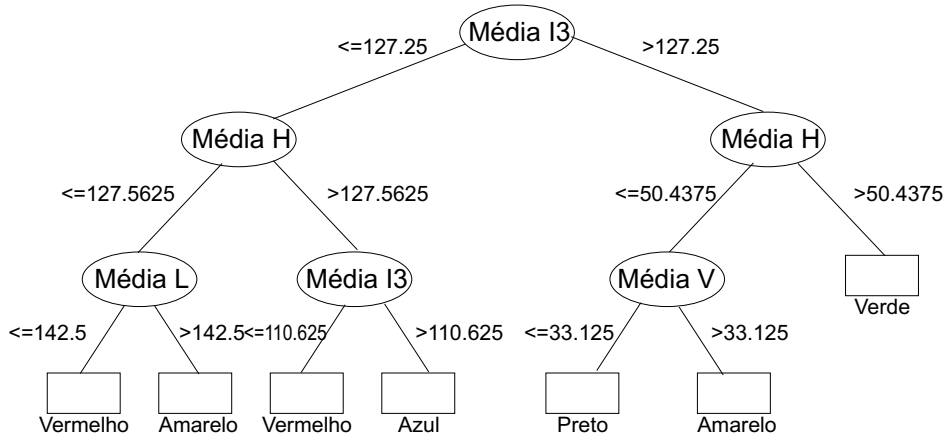


Fig. 9 - Árvore de decisão gerada

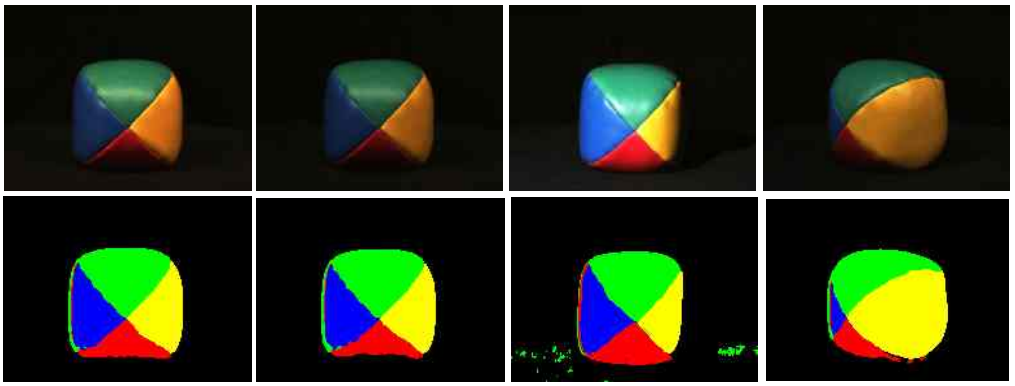


Fig. 10 - Imagens originais do objeto 25 da Biblioteca ALOI e os resultados da segmentação das imagens baseadas no método Wrapper com busca exaustiva.

No segundo experimento foi aplicada a metodologia proposta às imagens aéreas de citrus. O método *Wrapper* foi utilizado também em duas etapas, ou seja, somente com as médias em primeira instância e, posteriormente, adicionando-se as variâncias e entropias correspondentes às componentes selecionadas. A precisão em validação cruzada foi de 99.52% de acertos, sendo as componentes selecionadas as médias de a e de G, e a entropia de B, como mostrado na Figura 11. Os resultados da segmentação com a árvore da decisão (Fig. 11) são mostrados na Figura 12.

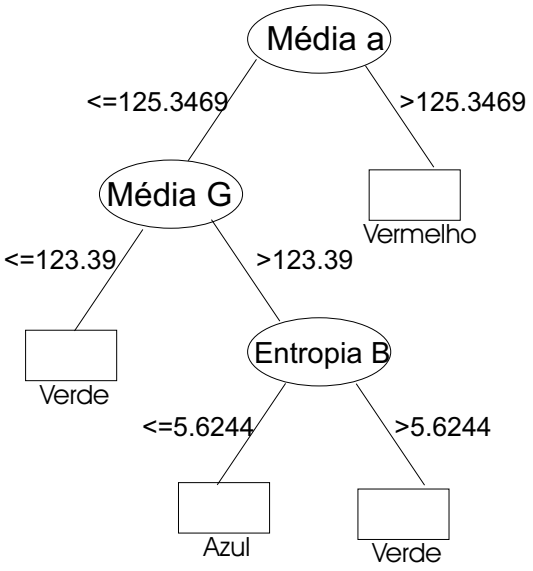


Fig. 11 - Árvore de decisão gerada para as imagens aéreas

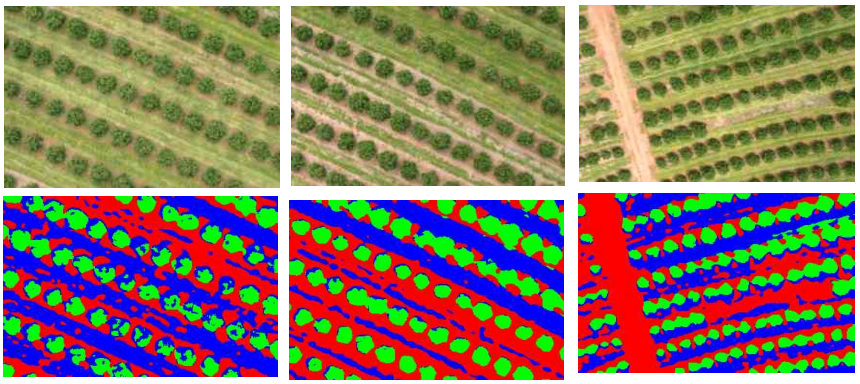


Fig. 12 - Imagens aéreas e os resultados da segmentação das imagens baseadas no método Wrapper com busca exaustiva.

O método *Wrapper* foi proposto e aplicado de modo a selecionar um subconjunto ótimo de componentes de modelos de cores para uma discriminatória e robusta segmentação de imagens. Foi verificado experimentalmente que o método *Wrapper* usando uma busca exaustiva na seleção do subconjunto de características com Árvore de Decisão pelo algoritmo C4.5, representa uma abordagem com bons resultados de segmentação. A seleção do subconjunto de componentes de cor efetuada permitiu o balanço apropriado entre repetibilidade e o poder discriminatório.

Em Visão Computacional existe um crescente interesse na seleção de características, onde várias questões ainda permanecem abertas. Os métodos de Mineração de Dados para a seleção de características têm sido propostos com algum sucesso. Muitos exemplos destas aproximações são focalizados em clusterização numérica, não havendo nenhuma evidência teórica ou experimental relacionada a seu comportamento em imagens coloridas. Este é um dos primeiros trabalhos neste sentido. As experiências conduzidas em uma grande variedade de imagens tanto artificiais quanto reais, mostraram que o método proposto é aplicável e com resultados significativos na segmentação de imagens coloridas em diferentes condições de iluminação.

Referências

BARANAUSKAS, J. A. **Extração automática de conhecimento por múltiplos indutores**. 2001. 181 f. Tese (Doutorado em Ciências da Computação e Matemática Computacional) Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.

BLUM, A. L.; LANGLEY, P. Selection of relevant features and examples in machine learning. **Artificial Intelligence**, Amsterdam, v. 97, n.1, p. 245-271, 1997.

GEVERS, T. Color in image search engines. In: University of Amsterdam. **Principles of Visual Information Retrieval.** London: Spring-Verlag, 2001.

JORGE, L. A.C.; RUIZ, H. S.; FERREIRA, E. J.; GONZAGA, A. Wrapper Approach to Select a Subset of Color Components for Image Segmentation with Photometric Variations. In: PROCEEDINGS of SIBGRAPI 2007. Brazilian Symposium on Computer Graphics and Image Processing, 20., 2007, Minas Gerais, Brazil. [S. l.: s. n.], 2007. p. 245-252. ISSN: 1530-1834. ISBN: 978-0-7695-2996-7

KOHAVI, R.; JOHN, G.. Wrappers for feature subset selection. **Artificial Intelligence**, Amsterdam, v. 97, n. 1-2, p. 273-324, 1997.

LANGLEY, P.; IBA, W. Average-case analysis of a nearest neighbor algorithm. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 13., Chambery, 1993. **Proceedings**... Chambery: [s.n.], 1993. p. 889-894.

LEE, H. D.; MONARD, M. C.; BARANAUSKAS, J. A. **Empirical comparison of wrapper and filter approaches for feature subset selection**. Referências 112. [São Carlos]: USP-ICMC, 1999. 46 p. (USP-ICMC. Relatórios técnicos, 94). ISSN: 0103-2569).

QUINLAN, J. R. **C4.5**: programs for machine learning. San Mateo: Morgan Kaufmann, 1993. 302 p. ISBN: 1-55860-238-0.

YANG, J.; HONAVAR, V. Feature subset selection using a genetic algorithm. **IEEE Intelligent Systems**, New York, v. 13, n. 2, p. 44-49, 1998.

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Instrumentação Agropecuária
Ministério da Agricultura, Pecuária e Abastecimento
Rua XV de Novembro, 1452 - Caixa Postal 741 - CEP 13560-970 - São Carlos - SP
Telefone: (16) 3374 2477 - Fax: (16) 3372 5958
www.cnpdia.embrapa.br - sac@cnpdia.embrapa.br

Ministério da Agricultura, Pecuária e Abastecimento

UM PAÍS DE TODOS
GOVERNO FEDERAL