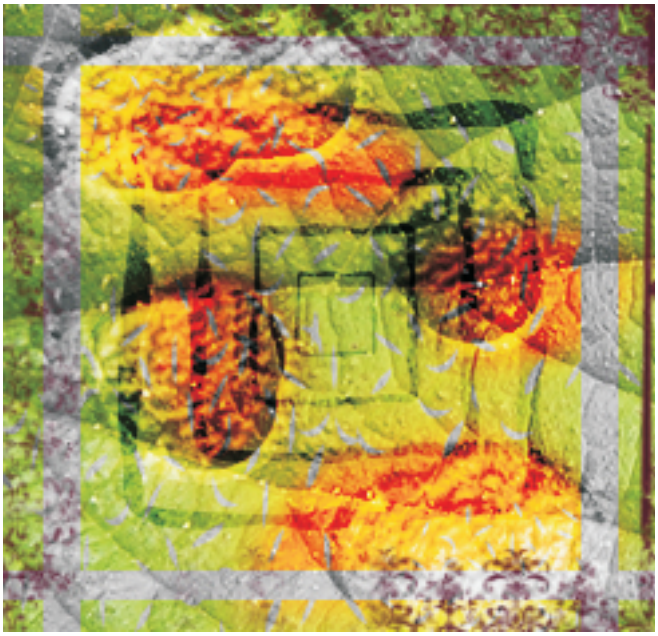
Comunicado Técnico 87

Dezembro, 2008
Campinas, SP



ISSN 1677-9274

LIVIA – um software para classificação não supervisionada de áreas foliares infectadas pela ferrugem do café

José Iguelmar Miranda[1]
João Camargo Neto[2]

O objetivo deste comunicado é apresentar a implementação Java™ do software LIVIA (Library for Visual Image Analysis). Trata-se de um módulo de processamento de imagens digitais aplicado à agricultura, desenvolvido na Embrapa Informática Agropecuária (Campinas/SP), sob demanda da Embrapa Meio Ambiente (Jaguariúna/SP).

O problema a ser solucionado era mensurar a extensão das áreas foliares infectadas pela ferrugem do café (Hemileia vastatrix) (Fig. 1). O trabalho de reconhecimento ou mensuração dessas áreas estava sendo feito visualmente, comparando as áreas infectadas na amostra foliar com um determinado padrão circular. O analista atribuía um valor, em porcentagem, da representatividade da infestação. Esse processo é altamente subjetivo e, até certo ponto, sujeito a erros por fatores como fadiga e imprecisão na acuidade visual do analista.

O problema apresentado se enquadra no procedimento clássico de reconhecimento de padrões, quando feições, as áreas de ferrugem, compartilham um conjunto de propriedades, neste caso, espectrais, semelhantes no espaço. No caso, a cena ou amostra é composta por imagens coloridas, portanto, estendem-se por três frequências visíveis do espectro eletromagnético, no azul, no verde e no vermelho.

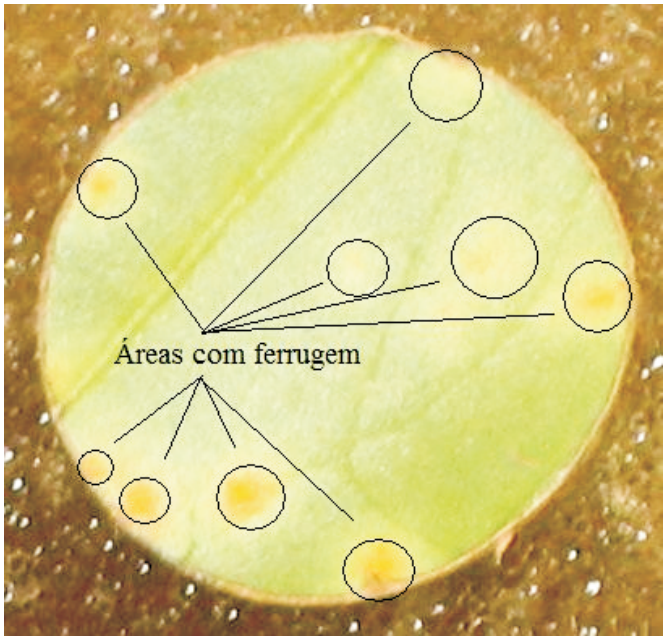


Fig. 1. Amostra de folha infectada com ferrugem do café.

[1] Ph.D. em Geoprocessamento, Analista da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: miranda@cnptia.embrapa.br)
[2] Ph.D. em Processamento de Imagens, Analista da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: camargo@cnptia.embrapa.br)

Embrapa

Existem dois procedimentos para o processo de classificar padrões: supervisionado e não supervisionado. No primeiro caso, o analista identifica os diferentes padrões ou categorias existentes na cena através de áreas de treinamento. Essas áreas consistem em um conjunto de medidas estatísticas que são usadas posteriormente por um determinado algoritmo que generaliza, para toda cena, os padrões identificados nas áreas de treinamento.

O segundo procedimento de reconhecimento de padrões ou classificação é não supervisionado. Nesse caso, todo o trabalho de identificação é de responsabilidade de um algoritmo que analisa toda a cena e identifica, com base em informações estatísticas, uma quantidade de categorias previamente definida pelo analista. A avaliação desses padrões estatísticos, permitindo uma classificação, é possível por causa do espaço de medidas ou espaço de atributos (Fig. 2). Todas as observações físicas descrevendo os padrões da cena estão contidas no vetor de medidas ou atributos x, que corresponde a um ponto nesse espaço p-dimensional, sendo p o número de atributos (dimensões) feitos. No presente caso, p = 3, as três cores básicas. Padrões diferentes, portanto, com diferentes valores de atributos, correspondem a diferentes pontos no espaço de atributos, já os padrões similares tendem a se agrupar em torno de pontos próximos no referido espaço.

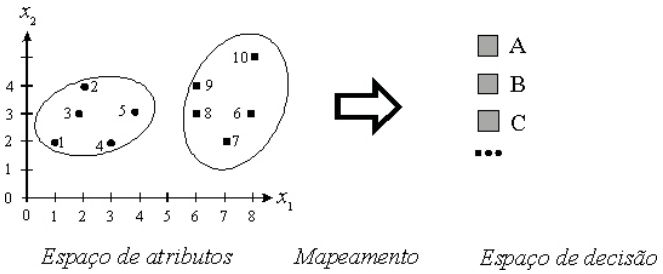


Fig. 2. Mapeamento do espaço de atributos no espaço de decisão.

A ilustração citada exemplifica um espaço de atributos bidimensional, com as variáveis x_1 e x_2. Nesse espaço, percebe-se a presença de dois padrões bem definidos que serão mapeados em uma das categorias do espaço de decisão (A, B, C, etc.). Como exemplo de um vetor de atributos, o ponto 9, seria x_9 = [6, 4], que corresponde às suas coordenadas no espaço de atributos, sendo 6 medidas na variável x_1 e 4 na variável x_2. O padrão pode ser definido pela dupla [x, δ], com x descrevendo seus atributos e δ, o seu significado ou descrição. O objetivo do mapeamento é estabelecer um relacionamento do espaço de atributos no espaço de decisão, permitindo rotular os pontos do espaço em suas respectivas categorias:

$$f: \mathbf{x} \rightarrow \delta$$

Sendo k o número de categorias a serem rotuladas, teremos no espaço de atributos k distribuições estatísticas específicas a essas categorias.

No processo de classificação, seja ele supervisionado ou não supervisionado, sempre haverá problemas de rotular alguns pontos do espaço de atributos, pois as variâncias das suas distribuições estatísticas nem sempre são concentradas, como seriam no exemplo mostrado na Figura 2. O mais provável é que existam áreas de sobreposição em maior ou menor intensidade. Nesses casos, haverá o problema de mistura de classes ou categorias. Portanto, o algoritmo implementado para realizar a classificação de áreas com a ferrugem terá esse problema, principalmente em áreas onde a concentração da doença seja de menos intensidade. Áreas com grande intensidade serão classificadas sem problemas.

As duas metodologias de classificação foram testadas, mas optou-se pelo método não supervisionado, visto que apenas uma categoria, ferrugem, precisava ser identificada. O desempenho do algoritmo k-médias, nesse caso, foi tão eficiente quanto os algoritmos supervisionados com menos dificuldades.

O algoritmo de classificação não supervisionada k-médias

Técnicas de agrupamento não hierárquico, como é o caso do k-médias, são feitas para agrupar objetos ou padrões em uma coleção de k grupos. A quantidade de grupos é definida antecipadamente. Na análise de agrupamento, os padrões são representados por pontos no espaço Euclidiano p-dimensional, onde os elementos dos vetores são valores dos atributos dos padrões e o objetivo é classificar os n padrões em k grupos, de maneira que uma certa medida de similaridade seja otimizada. A medida de similaridade é uma função de distância entre pontos dos padrões e seus centros, com o objetivo de minimizar essa função. Vários algoritmos existem para o problema de agrupamento, incluindo o k-médias, simulação de “annealing”, busca tabu e algoritmo genético. O primeiro algoritmo é uma técnica de busca local, enquanto os outros três são algoritmos de otimização global (Al-Sultan & Khan, 1996). É possível encontrar na literatura, referência ao k-médias como c-médias, por exemplo, Al-Sultan & Khan (1996) e Wu & Yang (2002), embora o termo c-médias seja mais usado em associação com o método difuso fuzzy c-médias.

Pode-se iniciar o processo de agrupamento do k-médias com uma partição inicial dos padrões em grupos ou um conjunto inicial de padrões “sementes,” que definirão os núcleos dos grupos. Uma maneira de iniciar é selecionar aleatoriamente pontos entre os padrões ou aleatoriamente particionar os padrões em grupos inicias. MacQueen, citado por Johnson & Wichern (1982), sugeriu o termo k-means (k-médias) para descrever seu algoritmo que atribuía cada padrão para o grupo tendo o centróide (média) mais próximo.

O problema do agrupamento pode ser assim resumido: dados n padrões em , alocar cada padrão em um, entre k grupos, minimizando a função critério, ou seja, a soma das distâncias quadráticas Euclidianas entre cada objeto e o centro do grupo (a ser encontrado), ao qual ele é alocado, deve ser mínima. O problema do agrupamento (PA), de acordo com Al-Sultan & Khan (1996), pode ser descrito matematicamente como:

minimizar
$$J(w, z) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} \left\| x_i - z_j \right\|^2 \quad (1)$$

tal que
$$\sum_{j=1}^{k} w_{ij} = 1 \text{ , } i = 1, 2, \ldots, n \text{ e } \leq$$

$$w_{ij} = 0 \text{ ou } 1, i = 1, 2, \ldots, n \text{ e } j = 1, 2, \ldots, k$$

em que k: número de grupos (definido), n: número de padrões disponíveis (definido), x_i Є : localização do i-ésimo padrão (definido), i Є [1, 2, ..., n], z_j Є : j-ésimo centro do grupo (centróide) (a ser encontrado) Є [1, 2, ..., k], w_{ij}: peso do padrão x_i com grupo j (a ser encontrado), dado por: 1, se o padrão for alocado ao grupo j; i = 1, 2, ..., n e j = 1, 2, ..., k, 0, em caso contrário, ∀: uma matriz n x k cuja coluna j é z_j como definido acima, w = [w_{ij}]: uma matriz n x k, $||x_i - z_j||^2$: distância Euclidiana quadrática entre padrão x_i e o centro z_j do grupo j.

Nessa formulação, a distância Euclidiana representa a medida de dissimilaridade. O PA é não convexo, possuindo muitos mínimos locais. A maior dificuldade é que a solução do PA pode levar a um ótimo local, mas não necessariamente global. O primeiro algoritmo usado para resolver o problema foi o k-médias, de Forgy (1965). Ainda hoje é um dos métodos mais usado para resolver o problema. Entretanto, ele pode falhar ao tentar a solução global do problema. Ball & Hall (1964) desenvolveram o algoritmo isodata para resolver o problema de agrupamento. O isodata é usado em alguns sistemas de processamento de imagens, mas cabe ressaltar que ele falha em achar uma solução ótima global, sendo, portanto, similar ao k-médias. Adicionalmente, ele é computacionalmente mais complexo e caro de implementar (Al-Sultan & Khan, 1996).

De acordo com Jain et al. (1999) e Xu & Wunsch II (2005), o algoritmo k-médias pode ser anunciado sucintamente da seguinte maneira:

1. Considere um conjunto de n padrões (vetores de atributos – Fig. 2) a serem agrupados. Inicializar aleatoriamente ou com base em algum conhecimento prévio, k partições (centróides de grupos) para coincidir com k objetos (2 k < n); ou k pontos definidos aleatoriamente dentro do hipervolume contendo o conjunto de objetos;
2. Atribuir cada padrão ao grupo com centro mais próximo: x_i Є c_w, se $||x_i – m_w|| < ||x_i – m_j||$ para todo i = 1, ..., n, j ≠ w, e j = 1, ..., k;
3. Recalcular os centros dos grupos usando a pertinência atual do grupo;
4. Se um critério de convergência não for satisfeito, voltar ao passo 2. Critérios típicos de convergência são: nenhuma ou mínima re-atribuição de padrões para novos grupos, ou que a função objetivo seja menor do que uma dada tolerância.

Aspectos da implementação

O LIVIA consta de sete programas: LiviaGUI, Amostra, Cluster, FiltrosImagem, Imagem, ImagemPainel e PainelControle. O LiviaGUI é o programa principal, responsável pela criação da interface gráfica (Fig. 3). Existem seis principais módulos: barra de menus, barra de ferramentas, janela de imagens, janela de resultados, painel de edição e painel de controle. A barra de menus permite ao usuário selecionar a imagem a ser classificada (Arquivo → Abrir). A opção "Processar" está desabilitada, permitindo futuras funções. A barra de ferramentas é um atalho rápido para as funções de abrir imagem e sair do programa. Na "janela principal" é mostrada a imagem selecionada para classificação. Depois que a imagem aparece nessa janela, o usuário clica em três pontos ao redor do disco de amostra foliar e o programa desenha um círculo com vinte pontos, pintados de verde (Fig. 4). Esses pontos podem ser movidos livremente no caso de se querer ajustá-los ou se a região para classificação não tiver a forma de círculo. No "painel de edição", as setas em "Mover" permitem a movimentação da região de interesse. Após definida a região de classificação, escolher o botão "Recortar". A região selecionada será mostrada na "janela de resultados" (Fig. 4).

O próximo passo é selecionar a quantidade de grupos desejados, no "painel de controle". Um máximo de sete grupos é permitido. Esse número pode ser modificado editando o programa PainelControle: private int maxGrupos = 7.

Depois de definida a quantidade de grupos, selecionar o botão "Classificar", que inicializa o algoritmo de classificação k-médias, conforme Jain et al. (1999) e Xu & Wunsch II (2005). A imagem recortada é desenhada com os grupos pintados em diferentes cores. Ainda no painel de controle, no lado direito ("Ferrugem"), existe um grupo de botões com as cores dos grupos. A idéia é que, por causa do problema do agrupamento, pode acontecer das regiões de ferrugem serem classificadas em duas ou mais cores diferentes, refletindo os diferentes estágios da doença. Nesses casos, o usuário pode selecionar aquelas cores que não correspondem à ferrugem.

Um exemplo é mostrado na Fig. 4. Escolheram-se seis grupos e depois que o usuário escolhe a opção "Classificar", o programa de classificação é executado e o resultado mostrado na "janela de resultados". Na

[4] Um reticulado é uma estrutura L = (L, R) tal que L é parcialmente ordenado por R e para cada dois elementos a, b de L existe supremo (menor limite superior) e ínfimo (maior limite inferior) de {a,b}.

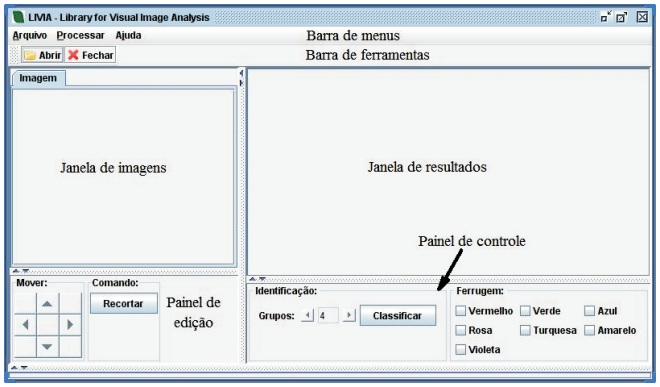


Fig. 3. Interface gráfica do LIVIA.

imagem classificada, a área "rosa" serve para o cálculo da área foliar com esporulação do patógeno e a área "verde" para calcular a área foliar lesionada. Notar que a área "verde" corresponde à área necrosada pelo fungo. A coloração "rosa", definida pelo classificador, equivale às estruturas reprodutivas do patógeno, que são de cor alaranjada na imagem original (daí o nome ferrugem). Além disso, o programa grava dois arquivos: um com a imagem classificada, no formato .png, e outro no formato .csv, formato que permite a leitura desse arquivo em qualquer planilha eletrônica (Excel, Calc, etc.). É importante frisar que esses arquivos são gravados no diretório "temp", que deverá ser criado no mesmo nível no qual o programa for instalado. Por exemplo, se o programa for instalado no diretório Windows: "c:\FerrugemCafe\LIVIA", então deverá ser criado o diretório "c:\FerrugemCafe\temp".

A lei de formação dos nomes desses arquivos é:

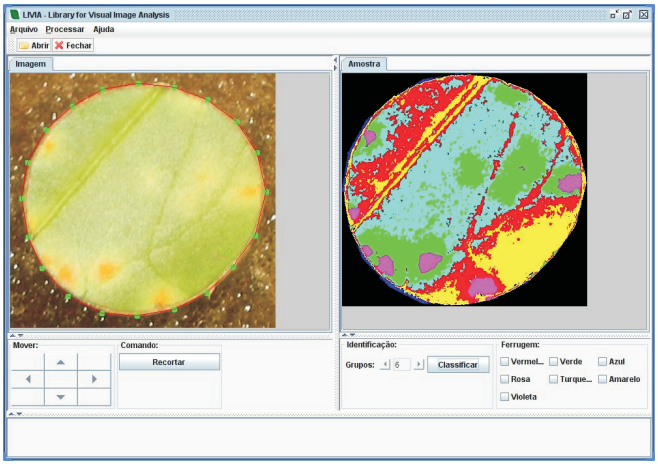


Fig. 4. Imagem e amostra classificada, com seis grupos.

<nome_do_arquivo>_<#grupos><kMed>. Considerando o arquivo "am_01.jpg", com seis grupos, o programa criará os arquivos "am_01_6kMed.png" e "am_01_6kMed.csv". A Tabela 1 apresenta um exemplo do arquivo gerado com a classificação da imagem mostrada na Figura 4 em seis grupos. Na primeira coluna, temos uma identificação genérica do número do grupo; na segunda coluna, a cor associada na imagem classificada com aquele grupo; na terceira coluna, a quantidade de pixels encontrada na imagem que corresponde àquele grupo; e, na quarta coluna, a porcentagem de pixels classificados no grupo, que corresponde, na realidade, à extensão, em termos percentuais, da área do grupo. A análise da imagem classificada na Figura 4, juntamente com os dados da tabela, permite concluir que 4,00% (grupo 4, rosa) correspondem à área foliar com esporulação do patógeno e que 18,00% (grupo 2, verde) correspondem à área foliar lesionada.

O LIVIA é distribuído livremente, sob a GNU General Public License (Licença Pública Geral) conforme

Tabela 1. Tabela 1. Amostra do arquivo .csv

Grupo	Cor	#pixels	%
Grupo 1	Vermelho	25151	24,00
Grupo 2	Verde	19125	18,00
Grupo 3	Azul	1608	1,00
Grupo 4	Rosa	4290	4,00
Grupo 5	Turquesa	35132	34,00
Grupo 6	Amarelo	17182	16,00

publicada pela Free Software Foundation, e se encontra disponível no diretório da rede agrolivre: (<http://repositorio.agrolivre.gov.br/projects/pid/>).

Conclusões

1. A utilização do algoritmo k-médias se mostrou eficiente na classificação das imagens, permitindo mensurar a área de dano da doença na amostra foliar;
2. A interface gráfica, mantida com um mínimo de recursos, se mostrou prática e útil na análise das imagens;
3. O programa está sendo testado com uma grande quantidade de amostras pelo Laboratório de Microbiologia Ambiental da Embrapa Meio Ambiente;
4. A definição de uma quantidade ótima de grupos para melhor classificar a imagem está também sendo realizada pelos pesquisadores daquela instituição;
5. A interface, desenvolvida com propósito específico, pode ser alterada para se adequar a novos tipos de análise

Referências Bibliográficas

AL-SULTAN, K. S.; KHAN, M. M. Computational experience on four algorithms for the hard clustering problem. *Pattern Recognition Letters*, v. 17, p. 295-308, 1996.

BALL, G. H.; HALL, D. J. Some fundamental concepts

and synthesis procedures for pattern recognition preprocessors. In: INTERNATIONAL CONFERENCE ON MICROWAVES, CIRCUIT THEORY, AND INFORMATION THEORY, 1964, Tokyo. *Proceedings…* [Tokyo: Institute of Electrical Communication Engineers of Japan, 1964].

FORGY, E. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, v. 21, p. 768, 1965. (Abstract).

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM Computing Surveys*, v. 31, n. 3, p. 264-323, 1999.

JOHNSON, R. A.; WICHERN, D. W. *Applied multivariate statistical analysis.* Englewood Cliffs, NJ: Prentice-Hall, 1982. 594 p.

WU, K-L.; YANG, M-S. Alternative c-means clustering algorithms. *Pattern Recognition*, v. 35, p. 2267-2278, 2002.

XU, R.; WUNSCH II, D. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, v. 16, n. 3, p. 645-678, 2005..

Comunicado Técnico, 87

Ministério da Agricultura, Pecuária e Abastecimento

BRASIL UM PAÍS DE TODOS GOVERNO FEDERAL

Embrapa Informática Agropecuária
Área de Comunicação e Negócios (ACN)
Endereço: Caixa Postal 6041 - Barão Geraldo
13083-970 - Campinas, SP
Fone: (19) 3211-5743
Fax: (19) 3211-5754
URL: http://www.cnptia.embrapa.br
e-mail: sac@cnptia.embrapa.com.br

1ª edição on-line - 2008