# Aplicação de técnicas de seleção e classificação de padrões na detecção de deficiência nutricial em espécies vegetais

Eduardo Pelli[1]
Heverton de Paula[1]
Fabricio Alves Silva[1]
Tamires Mousslech Andrade Penido[1]

[1] Universidade Federal dos Vales do Jequitinhonha e Mucuri – UFVJM
Campus JK Rodovia MGT 367 - Km 583, nº 5000, Alto da Jacuba – 39100-000 – Diamantina – MG, Brasil
pellie@ufvjm.edu.br, heverton.floresta@gmail.com, fabricioasv@gmail.com, penidotma@gmail.com

**Abstract.** The assessment of nutritional status of plants aims to diagnose which nutrients limit plant growth. The visual diagnosis associated with nutritional deficiency the standards previously established by symptoms on leaves. Through techniques of pattern recognition is possible to determine the most relevant features to classify a set of data. This work aimed the application of pattern recognition techniques in selecting the most relevant features for classification of nutritional deficiency analyzing significant regions of color histograms obtained from photographs of leaves. Three metrics for feature selection have been used to find the most relevant areas, namely: F-Score, Pearson correlation coefficient and Relieff for classification of data the technique of K-Nearest Neighbor (KNN) was used. It was possible to perform the classification of plants that possess or had normal nutritional deficiency obtaining an average accuracy of 79.2% is used as the most significant feature histogram of a small region of the red primary color. For future studies we intend to apply the techniques evaluated here for classification in classes covering the macro and micro nutrients.

**Keywords:** feature selection, classify, visual diagnosis Seleção de características, classificador, diagnose visual

## 1. Introdução

Reconhecer padroes faz parte da natureza humana, tarefas simples, como por exemplo identificar um estilo musical, reconhecer uma pessoa pela face ou identificar uma fruta pelo cheiro por meio dos sentidos da: audicao, visão, tato, olfato e paladar são possíveis devido a associação de características a objetos ou sensações. O reconhecimento de padroes consiste, portanto, na classificação ou categorização de um conjunto de dados por meio da análise de caracteristicas que os definem (BIANCHI, 2006).

A área de estudo denominada Visão Computacional busca desenvolver programas computacionais autônomos que consigam se assemalhar ao sistema visual humano, buscando analisar, interpretar e classificar imagens e objetos de interesse de forma confiável com base em padrões (PEDRINI; SCHWARTZ, 2008).

Uma máquina pode ser projetada e utilizada no reconhecimento padrões. A aplicação de uma máquina pode propiciar melhorias e evolução em vários processos como na identificação de impressões digitais, reconhecimento de sons e imagens aéreas, identificação de sequências de DNA dentre outras (DUDA; HART; STORK, 2001).

Os sistemas de visão computacional são constituidos por um conjunto de métodos e técnicas através dos quais sistemas computacionais podem ser capazes de interpretar imagens. Segundo Zúñiga (2012) um sistema de visão computacional é dependente do tipo e objetivo da aplicação.

No entanto, apesar da especificidade do sistema, a maioria segue uma serie de passos conforme ilustra a Figura 1.

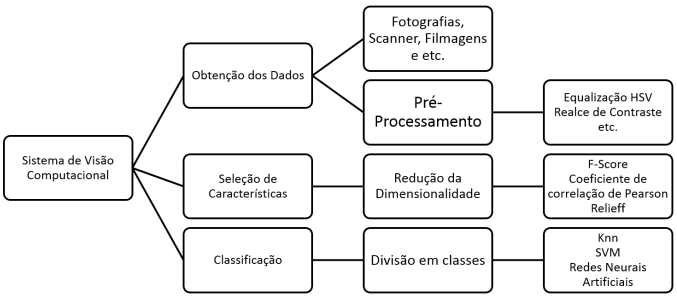


Figura 1: Etapas de Funcionamento de um sistema de visão computacional

Os problemas de seleção de variáveis podem ser classificados como supervisionados e não supervisionados, incluindo classificação, regressão, previsão de series temporais, clustering e etc (BI et al., 2003).

Na classificação não supervisionada o padrão é determinado por um limite de classe desconhecido, ou seja, caracteres são agrupados pelo algoritmo que busca a maior similaridade entre os pares. Já no aprendizado supervisionado o padrão do conjunto de dados é classificado ou categorizado previamente. Os padrões são especificados pela atribuição de valores para o conjunto de recursos e o objetivo é induzir que em novos casos seja possível chegar a novas classificações (BI et al., 2003; FORMAN, 2003) .

A escolha de sensores, técnicas de pré-processamento, esquema de representação e método para a tomada de decisão, depende do domínio do problema. Um problema bem definido e suficientemente detalhado, onde se tem pequenas variações intra-classes e grandes variações inter-classes, produzirá representações compactas de padrões e consequentemente a estratégia de tomada de decisão será simplificada. Aprender, a partir de um conjunto de exemplos (conjunto de treinamento), é um atributo importante desejado na maioria dos sistemas (BIANCHI, 2006).

Os nutrientes são elementos minerais, que têm fundamental importância no desenvolvimento vegetal, a ausência destes promove alterações no seu metabolismo gerando sintomas visiveis de deficiência. Em culturas comerciais é fundamental conhecer o nível nutricional das plantas, pois a ausência de nutrientes pode comprometer a produtividade da cultura, portanto, a avaliação do estado nutricional de espécies vegetais tem como objetivo identificar quais os nutrientes limitam o crescimento da planta, seja por excesso ou deficiência (MALAVOLTA, 2006).

As técnicas de avaliação nutricional basicamente consistem na comparação de padrões, a amostra analisada é comparada com indivíduos que possuem um padrão considerado ideal do ponto de vista nutricional. O método da diagnose visual associa as deficiências à padrões visuais (coloração, tamanho e forma) previamente estabelecidos (MALAVOLTA; VITTI; OLIVEIRA, 1997).

Este trabalho teve como objetivo a aplicação de técnicas de reconhecimento de padrões para identificação de deficiência nutricional em esécies vegetais por meio da análise de regiões mais significativas dentre os histogramas de cores obtidos das fotografias das folhas..

## 2. Material e Métodos

A base de dados do experimento foi obtida em condições naturais de campo em áreas do Parque Estadual do Biribiri no município de Diamantina/MG por meio de fotografias. Realizou-

se a classificação dos dados com base em aspectos visuais que dividiu o banco de dados obtido em duas classes, sendo: Folhas Normais (FN) e Folhas com Deficiência nutricional (FD). Foram obtidas 100 fotografias utilizando duas câmeras, sendo uma *Sony DSC HX1* com 9 megapixels e uma *Panassonic Lumix FZ-35* com 12 megapixels. A proporção de FD e FN foi de 50%, sendo que foram obtidas uma foto com deficiência e uma sem de cada espécie, escolhidas aleatoriamente. Foi utilizado um papel cartolina de cor preta como fundo das fotografias, visando mitigar possíveis variações de cores no plano de fundo, que poderiam gerar erro amostral na identificação da deficiência nutricional.

Para minimizar os efeitos da diferença de resolução entre as câmeras as imagens foram normalizadas dividindo-se os pixels de cada banda (RGB) pelo número total de pixels da imagem.

Os dados passaram por uma etapa de pré-processamento, onde onde foi aplicada a técnica Equalização de Histograma no espaço de cor HSV. Para a aplicação desta técnica as imagens são convertidas do RBG para o HSV, em seguida são obtidos os histogramas de luminância da imagem transformada, utilizando-se uma escala de mínimo e máximo do canal é aplicada a equalização, então os novos valores de luminância são salvos na imagem e a mesma volta a ser convertida para o espaço RGB.

Os dados foram aleatorizados e divididos em dois grupos, sendo um para treinamento com 60 fotos e um para teste contendo 40 fotos. A aleatorização e divisão da base de dados foi necessária para que a cada nova repetição os métodos de seleção de características utilizassem diferentes imagens para compor o grupo de treinamento e consequentemente o classificador utilizado apresentasse novos resultados.

Cada fotografia da base de dados foi convertida em três histogramas de cores com 256 classes de variações de tons, sendo um histograma para cada uma das cores primárias: vermelho (R), Verde (G) e azul (B) (Figura 2) .

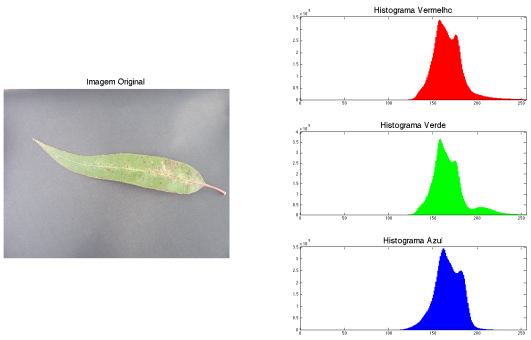


Figura 2: Histogramas de cores primárias

O grupo de treinamento foi obtido pela concatenação das matrizes oriundas dos histogramas de cores com dimensão $MatrizG_{60x256}$, $MatrizR_{60x256}$ e $MatrizB_{60x256}$ ,perfazendo uma matriz com dimensão $MatrizTreinamento_{60x768}$, sendo o índice 60 referente ao número de imagens, o 256 referente ao número de classes de variação de tons presente em cada histograma, e o 768 referente a soma do número de classes dos histogramas no RGB.

Foram aplicados três filtros de seleção, visando selecionar e ordenar as 10 características de maior importância para a solução deste problema. Os filtros utilizados foram: *F-score*, Coeficiente de correlação de Pearson e o *Relieff*

O *F-score* é um método simples para seleção de características. Este método é dado pela comparação direta das distâncias entre as médias de duas distribuições, em relação às suas

variâncias (DUDA; HART; STORK, 2001). A métrica *F-Score* é dada por:

$$f_i = \frac{(\bar{m}_i^{C_1} - \bar{m}_i) + (\bar{m}_i^{C_2} - \bar{m}_i)}{\sigma_i^{C_1} + \sigma_i^{C_2}} \quad (1)$$

Sendo que a característica $i$ que maximiza $f_i$ é referente à melhor separação entre duas distribuições $C_1$ e $C_2$, portanto, essa métrica pode ser utilizada como critério para seleção das características mais relevantes a serem aplicadas à um classificador.

O Coeficiente de correlação de Pearson mede o grau da correlação e a direção dessa correlação, podendo ser se positiva ou negativa, entre duas distribuições de características. O cálculo deste coeficiente é dado por (DUDA; HART; STORK, 2001):

$$\rho_j = \frac{\sum_{i=0}^{n}(x_{ij} - \bar{x}) \times (y_{ij} - \bar{y})}{\sqrt{\sigma_{xj} \times \sigma_y}} \quad (2)$$

Este coeficiente assume apenas valores entre -1 e 1. Quando $\rho = 1$: existe uma correlação perfeita positiva entre as duas características; Se $\rho = -1$, existe uma correlação negativa perfeita entre as duas características. Ou seja, quando o valor de uma aumenta, o da outra sempre diminui. Por fim, quando $\rho = 0$ é definido que as duas características não possuem dependência linear entre elas.

O *Relieff* é um método multivariado que diferentemente do *F-Score* e do coeficiente de correlação de Pearson, que apenas fazem a análise entre dois grupos, permite fazer análise multivariada e é um algoritmo que envolve parcialmente aprendizado de máquina (KIRA; RENDELL, 1992).

O *Relieff* realiza uma aproximações baseadas no KNN, na quais Calcula-se a ordem das características e pesos de atributos preditores para uma matriz de dados de entrada $X$ e $Y$, medindo a relevância individual de um atributo no contexto de outros . O algoritmo realiza uma ordenação dos atributos de acordo com um critério de importância, dado por pesos de atributos que variam de -1 a 1, com os maiores pesos positivos atribuídos a atributos mais importantes (ROBNIK-SIKONJA; KONONENKO, 1997).

O KNN diferentemente de muitos outros métodos de classificação não utiliza dados de treinamento para produzir um padrão de classificação. O processo de classificação do KNN é realizado de modo que, cada novo padrão a ser classificado, são realizadas buscas nos dados de treinamento com o intuito de verificar quais os dados mais se assemelham ao padrão que se deseja classificar, assim o objeto será identificado como pertencente à classe mais comum, ou seja, a classe que possui objetos mais similares a ele. A classificação é realizada por analogia, e não com a criação ou aplicação de algum modelo (DUDA; HART; STORK, 2001). Essa classe mais comum é proveniente da ideia de que objetos que se encontram mais "próximos"no conjunto de atributos têm mais possibilidades de pertencerem a uma mesma classe, ou seja, assume que as amostras correspondem a pontos em um espaço $n$-dimensional, onde $n$ é o número de descritores utilizados para representar as amostras. Desta forma, a classificação consiste em, para cada novo objeto determinar a classe dos objetos mais "próximos", para calcular a proximidade das amostras são utilizadas médidas de distância, podendo ser a distância euclidiana um exemplo (GUYON et al., 2006).

O valor de $k$ determina o número de vizinhos a ser considerado para a classificação, deve-se evitar valores muito pequenos para evitar que classificação fique sensível a pontos de ruído bem como números grandes para evitar incluir elementos de outras classes, é desejado que o $k$ seja também um número ímpar, para reduzir a ocorrência de impossibilidade de classificação devido a empate na quantidade de vizinhos mais próximos (SOUSA, 2013; DUDA; HART; STORK, 2001).

De posse das dez características discriminantes obtidas por meio dos filtros de seleção, foi apresentado ao classificador *K-Nearest Neighbor*(KNN) uma matriz com dimensoes $MatrizTreinamento_{60x10}$, sendo o índice 10 o número de características obtidas pelos filtros nas imagens do grupo de treinamento.

Em seguida aplicou-se o classificador sob o grupo de teste ($MatrizTeste_{40x10}$), sendo 40 o número de imagens do grupo de teste e 10 o número de número de características utilizadas, foram testados valore de K=1, K=5 e K=7.

Após a execução das 50 repetições, foi calculada a média e o desvio padrão dos 50 resultados obtidos com a execução do classificador para cada base pré-processada.

## 3. Resultados e Discussão

A técnica de Equalização de Histogramas *HSV* realizou modificações apenas na componente *V*, referente a luminância, e não alterou as componentes referentes a tonalidade de cores (*H* e *S*). Tal fato normalizou o histograma e permitiu aos filtros de seleção a escolha de novas características capazes de distinguir entre as classes observadas.

Na Figura 3 é apresentada uma imagem da base de dados utilizada, e seu respectivo histograma referente a sua componente *V* sem o pré-processamento:

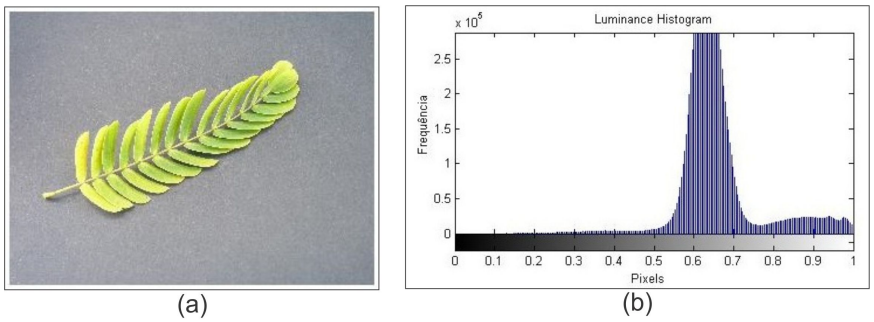


Figura 3: Exemplo de imagem original da base de dados (a) e seu respectivo histograma da componente *V* de luminância (b)

Após a aplicação da técnica de equalização de histogramas *HSV* sobre a imagem , obteve-se a nova imagem e seu respectivo histograma da componente *V* (Figura 4):

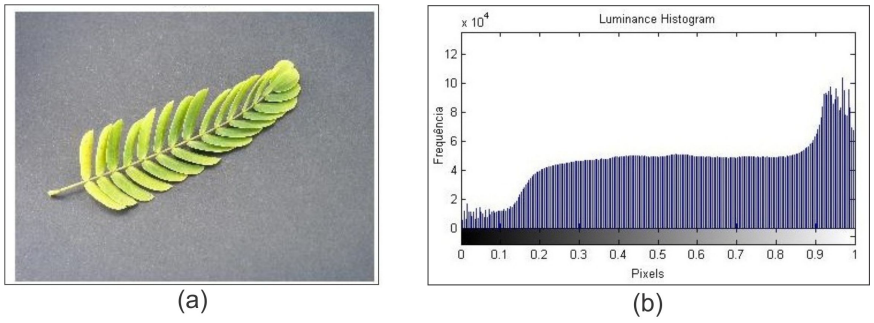


Figura 4: Exemplo de imagem pré-processada pela equalização de histogramas *HSV* (a) e seu respectivo histograma da componente *V* de luminância (b)

Como as componentes do modelo *HSV* referentes as componentes de tonalidade não foram modificadas com a aplicação da técnica de Equalização de Histogramas *HSV*, as alterações visíveis ao olho humano são mínimas.

Os três métodos de seleção aplicados apresentam na maioria dos cenários testados características presentes no intervalo de regiões $\{215\ e\ 230\}$, do histograma da cor vermelha, que podem ser consideradas determinantes na classificação de vegetais normais ou com deficiência nutricional.

Apesar de todos os métodos terem encontrado resultados semelhantes, é importante ressaltar que o filtro *Relieff* é capaz de realizar a busca de vários padrões de características discriminantes em mais de um grupo simultaneamente. Essa característica torna o *Relieff* um filtro muito interessante, pois o permite agrupar características para serem avaliadas conjuntamente como realizaram Plotze (2004) e Zúñiga (2012) que avaliaram características de textura combinadas a cor.

Já o *Coeficiente de correlação de Pearson* e o *F-Score* são capazes de analisar apenas um padrão de característica por vez, o que torna estes métodos limitados para a resolução de alguns tipos de problemas, porém pela sua simplicidade de uso convém testá-los.

Em termos práticos o *Relieff* faz a busca em uma $MatrizTreinamento_{(1x768)}$, sendo que o intervalo $(1 : 256)$ representa o vermelho, o intervalo $(256 : 512)$ representa o verde e o intervalo $(512 : 768)$ representa o azul, enquanto o *Coeficiente de correlação de Pearson* e o *F-Score* realizam a busca em uma $MatrizTreinamento_{(1x256)}$, que representa apenas o histograma vermelho.

Apos o classificador KNN ser aplicado a cada repetição com o valor de $K$ estabelecido sob o grupo de teste ($MatrizTeste_{40x10}$), foi obtida a acurácia média e o desvio padrão da classificação das classes propostas, na tabela (1) são apresentados estes valores assim como os valores máximo e mínimo da acurácia obtida para o conjunto de 50 repetições para cada valor de $K$ testado.

Tabela 1: Acurácia (ACC), desvio padrão (D.P), Valor Máximo (V. Max) e Valor Mínimo (V. Min) obtidos para diferentes valores de K

| | K=3 | | | | K=5 | | | | K=7 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | D.P | V. Max | V. Min | ACC | D.P | V. Max | V. Min | ACC | D.P | V. Max | V. Min |
| Relieff | 77,20 | 10,00 | 92,50 | 45,00 | 74,50 | 12,10 | 90,00 | 40,00 | 74,60 | 12,60 | 90,00 | 47,50 |
| Pearson | 79,20 | 5,60 | 92,50 | 62,50 | 79,30 | 5,90 | 90,00 | 55,00 | 79,60 | 6,70 | 95,00 | 52,50 |
| F-Score | 78,20 | 6,90 | 92,50 | 60,00 | 78,00 | 7,80 | 90,00 | 55,00 | 78,00 | 7,60 | 92,50 | 52,50 |

Para os cenários gerados a partir das características selecionadas pelo *Relieff* o valor de K=3 se mostrou mais interessante que os demais valores testados, (Figura 5).

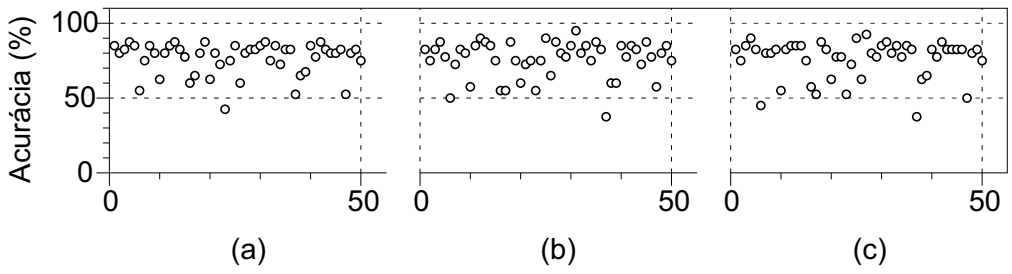


Figura 5: Distribuição da Acurácia de classificação nas repetições com características selecionadas pelo *Relieff* com diferentes valores de K. [a] K=3; [b] K=5; [c] K=7

O valor de K representa o número de vizinhos mais próximos a que o padrão selecionado será comparado, como o *Relieff* realiza a busca de características em um maior espaço existe

uma maior possibilidade do vizinho mais próximo estar distante e ter uma maior probabilidade de pertencer a outra classe, o que provoca erro na classificação.

Os cenários gerados a partir das características selecionadas pelo *Coeficiente de correlação de Pearson* apresentaram uma menor dispersão dos resultados de acurácia quando comparados aos obtidos pelo *Relieff*, conforme pode ser observado na Figuras(5 e 6)

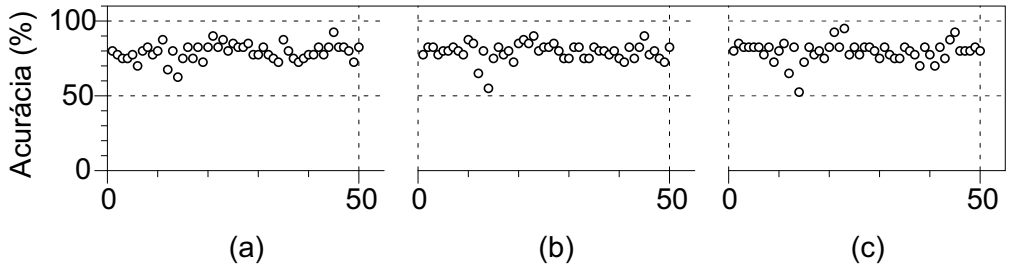


Figura 6: Distribuição da Acurácia de classificação nas repetições com características selecionadas pelo Coeficiente de Correlação de Pearson com diferentes valores de K. [a] K=3; [b] K=5; [c] K=7

Isto ocorre devido a este coeficiente apresentar uma pequena variação no intervalo de características selecionadas, o que pode conferir grande proximidade aos vizinhos mais próximos.

Os resultados obtidos utilizando-se as repetições geradas pelas características selecionadas pelo *F-Score* (Figura7) foram melhores quando comparados aos obtidos utilizando o *Relieff* (Figura 5), porém foram inferiores aos obtidos pelo *Coeficiente de correlação de Pearson* (Figura 6).

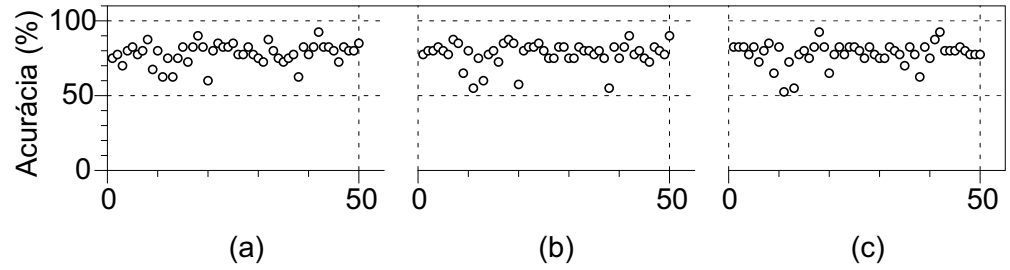


Figura 7: Distribuição da Acurácia de classificação nas repetições com características selecionadas pelo F-Score com diferentes valores de K. [a] K=3; [b] K=5; [c] K=7

Fato este ocorrido devido ao *F-Score* apresentar um maior intervalo $(141 : 228)$ de seleção para os 50 cenários que o Coeficiente de correlação de Pearson que apresenta um intervalo de $(185 : 228)$

O *F-Score* e o Coeficiente de correlação de Pearson apresentaram bons resultados para os diferentes valores de K testados.

Para a classificação de deficiências nutricionais em milho Zúñiga (2012) afirmou que o KNN é o classificador mais recomendado para utilização em técnicas baseadas na análise de cor, e utilizou K=1 alegando estar em busca de manter a simplicidade do modelo, abordagem esta que deve ser evitada para evitar que classificação fique sensível a pontos de ruído, haja vista que caso o espaço de busca seja considerado grande as classes podem não se aproximar.

## 4. Conclusão

Foi obtido um nível de $79,6\%$ de acerto na classificação de quais plantas possuíam deficiência nutricional, por meio da aplicação de técnicas do reconhecimento de padrões, na análise de fotografia de folhas.

A cor primária vermelha foi a mais significativa para se classificar as espécies vegetais com deficiência ou não, sendo que uma pequena região do histograma da cor vermelha às diferencia de maneira significativa.

Para estudos futuros podem ser analisados, por meio das técnicas utilizadas neste trabalho, a identificação das deficiências nutricionais específicas, avaliando as classes de micro e macro nutrientes, assim como avaliar a combinação de características de cor e textura, buscando uma maior taxa de acerto.

A metodologia proposta pode ser utilizada em outros problemas de classificação, bem como podem ser testados classificadores mais robustos, como o *Support Vector Machine* (SVM) e as Redes Neurais Artificiais.

**Referências**

BI, J. et al. Dimensionality reduction via sparse support vector machines. v. 3, 2003.

BIANCHI, M. F. *Extração de caracteríticas de imagens de faces humanas através de Wavelets, PCA e IMPCA*. Dissertação (Mestrado) — Escola de Engenharia de São Carlos - SP: Universidade de S?o Paulo, 2006.

DUDA, R. O.; HART, P. E.; STORK, D. G. Pattern classification. 2001.

FORMAN, G. An extensive empirical study of feature selections metrics for text classification. 2003.

GUYON, I. et al. *Feature Extraction - Foundations and Applications.* [S.l.: s.n.], 2006.

KIRA, K.; RENDELL, L. A. A practical approach to feature selection. In: MORGAN KAUFMANN PUBLISHERS INC. *Procedings of the Ninth International Workshop on Machine Learning*. [S.l.], 1992. (ML-92), p. 249–256.

MALAVOLTA, E. *Manual de Nutri??o de Plantas*. [S.l.]: Agron?mica Ceres, 2006.

MALAVOLTA, E.; VITTI, G.; OLIVEIRA, S. A. *Avaliação do estado nutricional das plantas: principios e aplicações*. [S.l.: s.n.], 1997. 319 p.

PEDRINI, H.; SCHWARTZ, W. R. *Análise de Imagens Digitais. Principios Algoritmos e Aplicações*. [S.l.: s.n.], 2008.

PLOTZE, R. de O. *Identificação de especies vegetais atraves da analise da forma interna de órgãos foliares*. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação - USP - São Carlos, 2004.

ROBNIK-SIKONJA, M.; KONONENKO, I. An adaptation of relief for attribute estimation in regression. 1997.

SOUSA, R. T. *Avaliação de classificadores na classificação de radiografias de tórax para diagnóstico de pneumonia infantil*. Dissertação (Mestrado) — Programa de Pós-Graduação do Instituto de Informática da Universidade Federal de Goiás, 2013.

ZúñIGA, A. M. G. *Sistema de Visão Artificial para identificação do estado nutricional de plantas*. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação - USP - São Carlos, 2012.